2016

Classifying and Predicting Disease Outcome from Continuous and Binary Predictors and Their Interactions

Sybil Prince Nelson
Medical University of South Carolina

Classifying and Predicting Disease Outcome from Continuous and Binary Predictors and Their Interactions

Sybil Prince Nelson

A dissertation submitted to the faculty of the Medical University of South Carolina
in fulfillment of the requirement for the degree of Doctor of Philosophy
in the College of Graduate Studies.

Department of Public Health Sciences

2016

Approved by:

Bethany Wolf, PhD — Chair

Viswanathan Ramakrishnan, PhD — Co-Chair

Paul Nietert, PhD

Diane Kamen, MD

Paula Ramos, PhD

Classifying and Predicting Disease Outcome from Continuous and Binary Predictors and Their Interactions

Sybil Prince Nelson

A dissertation submitted to the faculty of the Medical University of South Carolina
in fulfillment of the requirement for the degree of Doctor of Philosophy
in the College of Graduate Studies.

Department of Public Health Sciences

2016

Approved by:

Bethany Wolf, PhD Chair

Viswanathan Ramakrishnan, PhD Co-Chair

Paul Nietert, PhD

Diane Kamen, MD

Paula Ramos, PhD

ACKNOWLEDGMENTS

This presentation was made possible by grants from the National Institute of Arthritis and Musculoskeletal and Skin Diseases (P60 AR062755) and the National Institute of General Medical Sciences (T32GM074934) from the National Institutes of Health. SREB

TABLE OF CONTENTS

ABSTRACT

Many diseases have complex etiologies arising from interactions among genetic and environmental factors [1]. If an increased risk of disease is due to interactions between factors rather than a single factor alone, identification of the risk factors associated with the disease outcome can be difficult to detect using traditional statistical methods. For example, using a traditional logistic regression approach, interactions should be selected *a priori*, and sufficient data must be available in order to develop a model including interactions and their associated main effects. Also, if attempting to evaluate all possible interactions, the number of terms to include in logistic regression grows exponentially. In contrast, decision tree methods do not require identification of interactions *a priori*, and they can handle large numbers of variables. Classification and Regression Trees (CART) is a popular decision tree method, but it is biased toward the inclusion of continuous variables in the model [2]. It also can not exactly capture certain combinations of variables. Logic regression, an alternative decision tree method designed to find interactions among binary variables using Boolean logic, is able to identify exact interactions that are difficult to identify using CART. This dissertation extends logic regression methodology to allow for the inclusion of continuous variables within the logic regression framework. In order to do this, we first investigate which methods for dichotomization of a continuous variable to discriminate a binary outcome are the most effective for identifying the true threshold of a continuous variable, given one exists. Dichotomization methods are regularly used for patient risk stratification and in some statistical applications, for example to simplify interpretation of results; thus, it is important to know which dichotomization methods successfully identify a true threshold [3–9]. If the interaction of two or more variables, rather than their main effects lead to an increased risk of disease, then dichotomizing the variables in the interaction term individually may obscure the association with disease outcome, Y, making it more difficult to find the true thresholds. Thus, we also develop a method for jointly dichotomizing two or more variables to discriminate a dichotomous outcome in the case where the interaction between

the variables are associated with outcome, Y. We also use the dichotomization methods proven to theoretically recover a true threshold to develop an algorithm called C.Logic that allows for the inclusion of continuous variables in the logic regression framework.

Specific Aim 1

Evaluate the ability of different methods of dichotomizing a continuous variable to discriminate a binary outcome to recover the true threshold, given one exists.

a. Theoretically show which methods of dichotomization should recover a true threshold, T.

b. Compare the methods theoretically shown to identify a true threshold when sampling from a population. The performance of each method will be measured by examining the Mean Squared Error (MSE) and bias of the estimated threshold.

Specific Aim 2

Develop an algorithm for joint thresholding two or more continuous variables to discriminate a binary outcome.

a. Theoretically support the argument for using joint thresholding of interacting variables as opposed to single thresholding.

b. Compare joint thresholding to single thresholding when sampling from a population. The performance of each method will be measured by examining MSE and bias.

Specific Aim 3

Develop an algorithm, C.Logic, that allows for the inclusion of continuous variables and their interactions within the logic regression framework.

a. Compare the performance of C.Logic to CART for identifying continuous variables and their interactions that are associated with a binary outcome. The performance will be measured by how many times an interaction is exactly identified in the model.

INTRODUCTION AND SIGNIFICANCE

1.1 Introduction

Dichotomization of continuous variables is frequently used in clinical and statistical applications. Physicians may want to stratify patients according to risk, make determinations about the necessity of additional diagnostic testing, or allocate physician resources according to patient need. For example, a prostate-specific antigen (PSA) greater than 4 ng/mL is often used to determine whether or not a patient should undergo a biopsy to test for prostate cancer. Similarly, a total cholesterol level greater than 200 mg/dL indicates increased risk of heart disease and thus is useful to determine if a patient should receive cholesterol lowering medication [3, 4]. Additionally, statisticians may dichotomize continuous variables to improve interpretability of statistical models or for use in some statistical models such as decision trees that require that all variables be dichotomized prior to or during implementation [6–9]. If a true threshold exists that discriminates between two groups, the challenge is identifying which of the many methods available for dichotomization will recover it. A detailed discussion of dichotomization methods will follow in chapter 2.

Though dichotomization of continuous variables is common, there are some negative consequences associated with the practice such as loss of power and residual confounding [10]. Therefore, it is important to identify accurate methods of dichotomization to minimize these consequences while dichotomizing variables necessary for clinical or statistical applications.

A growing body of evidence suggests that complex diseases such as Alzheimer's disease and diabetes mellitus may result from interactions between multiple genetic and environmental factors as opposed to rare Mendelian diseases such as sickle cell anemia and Hutchinson-Gilford progeria syndrome which are characterized by gene variants in a single gene [1, 11, 12]. Dichotomization of continuous variables is further complicated when considering the effect of interactions of variables on an outcome. If continuous variables are associated with other variables to increase risk of disease only through their interaction and not their main effect, yet they are dichotomized independently,

the thresholds chosen may be suboptimal. Thus, there is a need to develop a method to dichotomize interacting continuous variables jointly as opposed to individually.

1.1.1 Clinical Motivation

In addition to Alzheimer's disease and diabetes mentioned above, systemic lupus erythematosus (SLE) is another complex disease hypothesized to develop and progress as a result of complex interactions among genetic and environmental factors. SLE is an autoimmune disease characterized by the production of autoantibodies against nuclear antigens. The incidence of SLE in the United States has tripled in the past four decades [13] and SLE has one of the highest death rates among rheumatic diseases [14].

In the U.S., the prevalence of SLE is estimated to be between 0.05% and 0.1% of the population and disproportionately affects females and African Americans [15]. African American(AA) men have prevalence of 0.038% compared to that of 0.009% in white men. The prevalence in AA women is 0.282% compared to that of 0.038% in white women [15]. AA women and men are also known to manifest more severe complications (e.g. central nervous system vasculitis, pulmonary hypertension, interstitial lung disease, stroke and death) and a younger onset of SLE [16]. Lupus nephritis (LN) is one of the most common and severe manifestations of SLE and is exhibited in approximately 50% of SLE patients [17]. It is more common and severe in AAs and more often leads to end-stage renal failure [17].

Studies support the genetic etiology of SLE [18, 19] as well as give evidence of possible environmental triggers such as ultraviolet light, infections, smoking and medications [20]. Therefore, there is a need to develop statistical methods for the identification of gene-gene and gene-environment interactions and their association with disease [21].

1.1.2 Statistical Motivation

Some statistical techniques used to model a dichotomous outcome given a set of factors include logistic regression, artificial neural networks (ANN), support vector machines (SVM), linear discriminant analysis (LDA), logic regression and decision trees (DT). Many of these methods can also

incorporate interactions in their model, but for one reason or another fall short of the goals for this dissertation. These methods are briefly described in the following sections.

Logistic Regression

Logistic regression (LR) is a statistical modeling approach that can be used to describe the relationship between $k \geq 1$ predictor variables $\mathbf{X}_i = (x_{i1}, x_{i2}, ..., x_{ik})$ and a dichotomous outcome variable ,y_i. Define P_{i1} as the probability subject i belongs to group 1, $P_{i1} = P(Y_i = 1|X_i)$, and P_{i0} as the probability subject i belongs to group 0 ,$P_{i0} = P(Y_i = 0|X_i)$. Then P_{i1}/P_{i0} is the odds of the i^{th} individual belonging in group 1. The natural logarithm of the odds ratio is equivalent to a linear function of the independent variables,

$$logit(P_{i1}) = ln(\frac{P_{i1}}{P_{i0}}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik},$$

The parameters $\boldsymbol{\beta}_j$ for $j = 0, ..., k$ of the logistic model are estimated using the maximum likelihood method. The probability $P(y_i = 1|\mathbf{x}_i)$ is estimated according to

$$P(y_i = 1|x_i) = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}}}$$

Although logistic regression can model the probability of disease as a function of predictor variables and their interactions, there are several disadvantages. First, if an interaction is included in the model the main effects also should be included. Thus, if it is believed that a factor increases risk of disease only in the presence of another factor and not with the main affect alone, the main effects must still be included in the model. Second, all possible variables should be identified *a priori*. If researchers are unsure as to what variables may or may not be associated with outcome, the model may have little predictive value if relevant independent variables are not used. These variables should be determined before use in logistic regression for optimal results. Third, including all possible interactions into a logistic model quickly becomes implausible as the number of terms to include grows exponentially. For example, if there are three variables, X_1, X_2, X_3, then there would be $2^3 - 1$=7 terms to include in the model. If there are 20 variables, the number of terms to include would be $2^{20} - 1$=$1,048,575$. Further, in order to produce optimal results, a rule of thumb

is that there should be at least 10 observations per predictor in the model which would also not be feasible [22].

Artificial Neural Networks

Artificial neural networks (ANN) are algorithms that can be used for nonlinear modeling. It is most commonly used as an alternative to LR for developing predictive models for a binary outcome in medicine [23]. In an advantage over LR, ANN has the ability to implicitly detect complex nonlinear relationships between outcome and predictor variables as well as detect all of the possible interactions between predictor variables [23]

Figure 1.1 is a diagram illustrating an ANN used to predict outcome based on two predictor variables or input variables, age and sex. The circles in this diagram are nodes. The lines are connection weights which can be compared to regression coefficients in the logistic regression setting [24]. Typically, an ANN has three layers, input, hidden, and output as noted in Figure 1.1. The calculations that occur in the hidden nodes allow the network to model nonlinear relationships between the predictor variables and the outcome.

One disadvantage of ANN is that it creates a black box effect making it difficult to understand the exact nature of the relationship between predictors and the outcome [23, 24]. Other disadvantages include increased computational burden and a tendency to overfit the data [23].

Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a technique used for data classification and dimensionality reduction. LDA optimizes the separation between classes by maximizing the ratio of between-class variance to the within-class variance of the data. To do this it creates a transformation of the data that provides a proper separation between the classes. It first takes a given data set and test set which are subsets of an original data set and finds the mean of each [25]. Thus there are three means μ_1, μ_2, and μ_3, where μ_3 is the mean of the entire data set. Next, the within-class and between-class scatter of the groups are used to formulate criteria to separate the classes [25]. The solution obtained by maximizing the ratio of between-class variance to within-class variance is the criterion that defines the axes of the transformed space which separates the classes [25].

Like ANNs, LDA does not specifically aid in finding interactions between predictors. Another

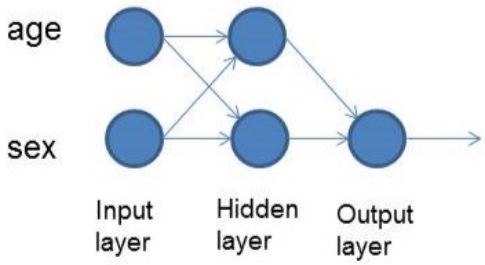

Figure 1.1: Sample Artificial Neural Network

problem with use of LDA to classify disease is that if discriminatory information is not in the mean of the predictors, analysis will fail. LDA is a parametric method that assumes Gaussian likelihoods, and it assumes the predictors are linearly separable which may not be the case.

Support Vector Machines

Support vector machines (SVMs) are another machine learning model that can be used for classification of a binary outcome. SVM algorithms develop a classifier by building a non-linear boundary (hyperplane) that provide the optimal separation of classes within a data set [24]

More formally, consider data D=$\{(\mathbf{x}_i, y_i) : i = 1, 2, . . .n\}$ where $\mathbf{x}_i \in R^r$ is a vector of

independent variables and $y_i \in \{0, 1\}$ is the outcome. A function f that will separate the observations does so by defining a hyperplane that is used to classify each new x. For instance, if there exists a linear separation of the data D , then a hyperplane could be written as a simple linear model,

$$\hat{y}(\mathbf{x}) = \beta_0 + \mathbf{x}^T\beta \tag{1.1}$$

where β is the weight vector and β_0 is the bias. In many instances, there are an infinite number of possible hyperplanes. The optimal separating hyperplane is one in which the distance from the hyperplane to the closest observation is maximized.

While SVMs do predict effectively, like ANNs they tend to overfit and they are computationally intensive. Also, no probability of class membership is given [24] so the results of SVMs are not easily interpreted.

1.2 Decision Trees

The methods described above do not focus on identifying interactions, but rather on prediction. With the exception of logistic regression, the relationship between interactions and the outcome is unclear. When important interactions are not known *a priori*, all $2^k - 1$ possible terms could be considered which quickly becomes computationally intensive. Also, considering that for each term in the model, there should be at least 10 observations for even relatively small values of k, such as k=20, it is unlikely that a large enough data set would be available to provide interpretable results. Decision trees are easily interpretable statistical models and can be represented as tree-like graphs or models describing how predictors are associated with outcome. A general decision tree consists of nodes which represent points where the tree splits based on a condition or rule. In general, the first rule in a decision tree best splits the entire data set into the most homogeneous subgroups with respect to a binary outcome. Then another rule is applied splitting the subgroups into further subgroups. This continues until a stopping rule is applied. The final subgroup groups are called leaves and consist of several variables where the splits occurred. These variables may represent interactions. Thus, in this dissertation, we focus on a decision tree approach which is an efficient way to identify possible

interactions associated with a dichotomous outcome.

1.2.1 Classification and Regression Trees (CART)

CART is a decision tree method that splits data into increasingly homogeneous groups based on characteristics of the predictor variables. It models response by recursively partitioning the training data and continuously selecting splits until a stopping criterion is implemented such as Gini impurity or entropy [26]. To prevent the overfitting that occurs in some other methods, the trees may be pruned (nodes deleted) after the stopping criterion is met. Pruning occurs by growing a tree until a minimum node size is reached and then deleting nodes back to an optimal size that is determined by cross-validation or using a test set [27]. A terminal node $m = 1, 2, ..., M$ of a CART tree corresponds to the final group (or region R_m)where an observation has been placed. If there are n number of observations in the data, then n_m represents the observations from the training data set whose predictors lead to node m. We can estimate the probability of being in a certain category by letting $\hat{p}_{ml}$ be that estimated probability from category l at node m. We are considering a binary outcome so the response classes for l are $l = 0, 1$. The probability of category l at node m is estimated by $\hat{p}_{ml} = \frac{1}{n_m} \sum y_g : \mathbf{x}_g \in R_m I(y_g = l)$ where the g^{th} response at node m is $y_g, g = 1, 2, ..., n_m$. Once the classes are defined using the training set, they are used to make predictions on new observations.

Figure 1.2 gives an example of a simple CART tree built from three binary predictors. This tree predicts an individual to be in category 0 if $X_1 = 0$ and $X_2 = 0$ or if $X_1 = 0$, $X_2 = 1$, and $X_3 = 0$. It predicts category 1 if $X_1 = 1$, or if $X_1 = 0$, $X_2 = 1$ and $X_3 = 1$.

While CART trees are easily interpreted and can identify complex interactions to model outcome, it is biased toward the inclusion of continuous variables into the model over binary [2]. Also, though CART is rather flexible, by design once a branch is formed, specific combinations of variables are potentially limited as can be seen in figure 1.3.

1.2.2 Logic Regression (LR)

Logic regression is an alternative decision tree method used when both the predictors and the outcome are binary. LR constructs Boolean combinations ("or"=$\wedge$, "and"=$\vee$, "not"=!) of the predictors in

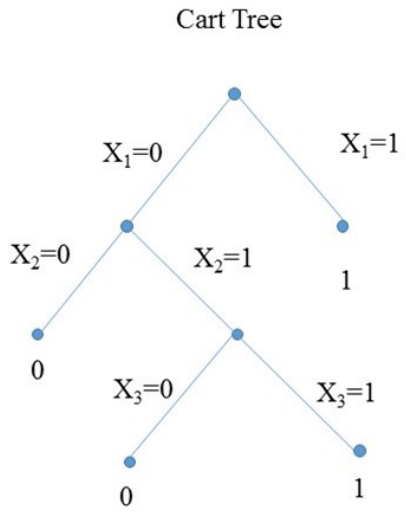

Figure 1.2: A CART tree with three binary predictors ($X_1, X_2, and X_3$) and two classes (0 and 1).

order to model an outcome. While LR can be used to model binary or continuous outcomes, our focus here is on LR for binary outcomes. Simulation studies show that LR is more efficient than CART at identifying complex interactions in data sets used to model the development of certain diseases [28].

The use of Boolean logic in the construction of decision trees provides LR with more flexibility than CART. This is because while all CART trees can be written using Boolean logic not all Boolean logic statements can be written as a CART tree. For example, suppose the interactions between variables (X_1 and X_2) or (X_1 and X_3) increase the risk of disease. We can write this interaction using the following Boolean statement: $(X_1 \vee X_2)$ or $(X_1 \vee X_3)$. As can be seen in Figure 1.3 this statement can only be modeled with a logic regression tree. The CART tree in Figure 1.3 predicts disease for variables X_1 and X_2 or for X_1, X_2, and X_3. In this instance, there is no way for CART

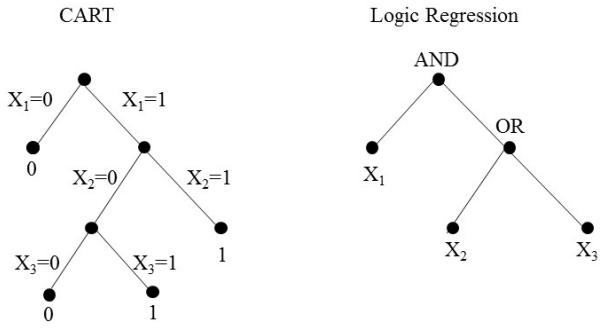

Figure 1.3: Sample CART V logic regression

to predict X_1 and X_3 leading to increased risk of disease.

Another advantage of LR is the use of simulated annealing in order to choose the logic regression models. In the LR setting, the entire set of predictors and predictor interactions is the state space where the collection of different configuration of the predictors represent the individual states. At initiation of the simulated annealing algorithm, an initial state is selected at random from among the possible states and a score is estimated based on some statistic (e.g. the misclassification rate for classification trees). The simulated annealing algorithm then randomly moves to a new state based on one of 6 permissible moves (i.e alternating a leaf, splitting a leaf, growing a leaf) and estimated a score for the updated state. If the score of the new state is better (i.e. smaller misclassification) than the previous state, the new state is accepted with probability = 1. If the new state is not better than that of the old state, the new state may still be accepted with a specific probability that determined by a parameter for the annealing algorithm referred to as the temperature. The annealing temperature is used to calculate the probability of accepting a worse state and decreases as the annealing chain progresses. As a result, the probability of excepting worse new states decreases as the annealing process [6] progresses. The initial temperature should be selected such that approximately 90-95% of

worse states are accepted and the final temperature should be selected such that fewer than 1-2% of worse models are accepted. The number of iterations selected for annealing process then determines the rate at which the annealing temperature decreases and is generally large (i.e. 100,000-250,000 iterations). The main advantage of simulated annealing is that is reduced the probability of selecting a logic regression trees that represents a local optima when searching through the state space of predictors unlike CART which uses a greedy search that is more likely to find a local optima.

Like CART, LR models are easily interpretable, but LR has the added flexibility introduced through Boolean logic. Consider the example of LR shown in Figure 1.4 where the predictors Age or Blood Pressure, or an interaction between single nucleotide polymorphisms (SNPs) 4 and 16 lead to increased risk of disease. Notice, however, that Age and Blood Pressure are inherently continuous variables. As LR is not currently designed for the inclusion of continuous variables, Age and BP would first have to be dichotomized. Considering that in the CART method the continuous variables are also dichotomized (at the split), this is not unusual. However, one must determine an effective method for dichotomizing the variables that best reveals the pattern in the data.

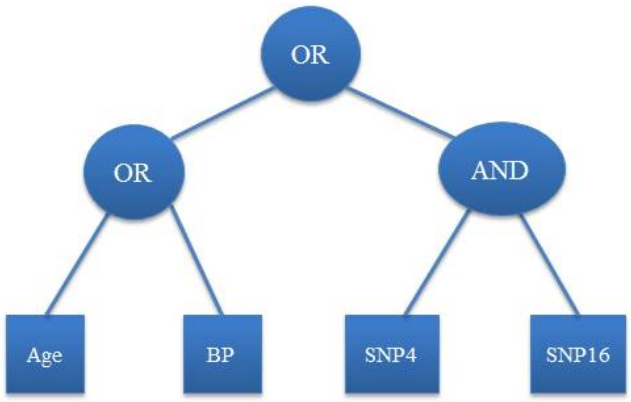

Figure 1.4: Sample Logic Regression Tree

1.3 Specific Aims

The goal of this dissertation is to develop statistical methodology that focuses on identifying interactions between genetic and environmental factors associated with increased risk of disease. To do this, we first identify which of several commonly used dichotomization methods are capable of correctly identifying a true threshold for a continuous variable to discriminate a binary outcome, given one exists. Next we develop a method for jointly thresholding two continuous variables that consider their interaction when dichotomizing. We conduct simulation studies, controlling the value of the true threshold of the continuous variables, the strength of association with binary outcome to evaluate which dichotomization methods best recover a threshold. We use this joint thresholding method in an algorithm called C.Logic to allow for the inclusion of continuous variables in the logic regression framework. Finally, we apply C.Logic to data from a study of subjects with lupus nephritis (LN) in order to identify interactions between predictors that are associated with treatment response in LN.

Specific Aim 1

Evaluate the ability of different methods of dichotomizing a continuous variable to discriminate a binary outcome to recover the true threshold, given one exists.

a. Theoretically show which methods of dichotomization should recover a true threshold, T.

b. Compare the methods theoretically shown to identify a true threshold when sampling from a population. The performance of each method will be measured by examining the Mean Squared Error (MSE) and bias of the estimated threshold.

Specific Aim 2

Develop an algorithm for joint thresholding two or more continuous variables to discriminate a binary outcome.

a. Theoretically support the argument for using joint thresholding of interacting variables as opposed to single thresholding.

b. Compare joint thresholding to single thresholding when sampling from a population. The performance of each method will be measured by examining MSE and bias.

Specific Aim 3

Develop an algorithm, C.Logic, that allows for the inclusion of continuous variables and their interactions within the logic regression framework.

a. Compare the performance of C.Logic to CART for identifying continuous variables and their interactions that are associated with a binary outcome. The performance will be measured by how many times an interaction is exactly identified in the model.

AN EVALUATION OF POPULAR DICHOTOMIZATION METHODS

2.1 Introduction

Dichotomization of continuous variables is frequently used in medical applications to stratify patients according to risk, make determinations about the necessity of additional diagnostic testing, and to allocate physician resources according to patient need [5, 29, 30]. For example, elevated low-density lipoprotein cholesterol (LDL-C) is a known cardiovascular disease risk factor. Determining a risk-benefit threshold LDL-C level of ≥ 190 mg/dL was instrumental in the development of guidelines for initiating statin therapy for primary prevention of cardiovascular disease [4, 31]. Additionally, in statistical modeling, dichotomizing continuous variables often results in a simpler interpretation, and some statistical models, such as decision tree methods, require that all variables be dichotomized prior to or during implementation [6–9, 32, 33].

If a true threshold exists that discriminates between two groups, the challenge is identifying it. Many methods are available for dichotomizing continuous predictors to discriminate between two groups. Some of these methods are based on expert opinion and epidemiological studies, such as with cholesterol level [4, 34, 35]. There are also many data-driven methods that select a threshold based on maximizing or minimizing a specific statistic. For example, the threshold could be chosen such that it maximizes the odds ratio between dichotomized predictor and the dichotomous outcome.

Although the primary purpose of all methods is to find a dichotomy that effectively discriminates between two groups, because the dichotomy is defined using a threshold, the problem reduces to effectively finding that threshold. This paper examines which of the most commonly used methods for dichotomization effectively recover a “true" threshold given that one exists. Section 2 of this paper defines the statistics to be maximized for dichotomization in terms of 2x2 contingency tables. Additionally, in this section, mathematical and numerical proofs regarding which of the methods recover the true threshold are also provided. Section 3 describes a simulation study that evaluates the impacts of location of the threshold, sample size, and strength of association between a continuous

predictor and a dichotomous outcome on the ability to recover the true threshold to provide guidance on which statistics are most effective and when these statistics are likely to fail. Section 4 presents the results of the simulation study. Section 5 provides a discussion of the implication of the results and offers recommendations regarding the appropriateness of a method of dichotomization for different scenarios.

2.2 Criteria for Dichotomization

Methods that can be used for dichotomizing a continuous predictor to discriminate between two groups can be separated into three main categories. Methods in the first category are clinically motivated using prior knowledge or experience [36–42] and are not supported by statistical theory. A second category of methods used for dichotomization is based on the prevalence of a condition in a population, such as observed prevalence which chooses a threshold, t, closest to the observed prevalence (i.e. $\frac{t}{max_t\|t-p\|}$ where p is the prevalence) [43,44]. Although methods based on prevalence are data-driven, the observed prevalence in the sample is dependent on the selected sample and may not reflect the population level disease prevalence. For example, in a 1:1 case-control scenario, the observed prevalence is determined by the study design rather than the natural prevalence in the population, in which case these methods will fail [44, 45]. Thus, methods based on prevalence, such as mean prevalence, matching prevalence, and observed prevalence, are not considered in this paper. Methods in the third category, the main focus of this paper, are data driven algorithms where the choice of threshold is selected by maximizing or minimizing a statistic, specifically Youden's statistic [46], odds ratio [47], ROC curve [48–50], relative risk [48], Gini Index [51], sensitivity and specificity [52] among others [32, 38, 53, 54]. Relative risk is only considered in the cohort study design where the sample is designed to mimic disease distribution in the population.

2.2.1 Numerical evaluation of the "best" threshold

This section provides an empirical examination of which of the common methods of dichotomization correctly identifies the true threshold, T. For the numerical investigation, the threshold, T, is specified

such that for a continuous random variable X and a binary random variable Y, $P(Y = 1|X \geq T) > P(Y = 1|X < T)$. We set $P(X > T) = 0.05$, $P(Y = 1) = 0.1$, and $P(Y = 1|X > t) = 0.4$. Then X is dichotomized for varying thresholds in a specified interval. For simplicity, we let $X \sim N(0, 1)$ and consider the range of X from $[-4, 4]$ in the increments of 0.001 with the true threshold included in the interval. Each value in $[-4, 4]$ is considered as a possible threshold for dichotomizing X to discriminate values of Y. For each possible threshold, t_x, there is a corresponding 2x2 contingency table between the dichotomized random variable X and Y with cell probabilities a, b, c, d as shown in Table 2.1. All statistics considered in this paper are defined in Table 2.2 using a, b, c, d and $P(Y = 1)$. Using the 2x2 table, the statistics defined in Table 2.2 are calculated using the probabilities depicted in Figure 2.1. The cell probabilities for $t_x < T$, $t_x = T$, and $t_x > T$ are shown below.

1. $t_x < T$

$$\begin{aligned} a &= P(X \geq T)P(Y = 1|X \geq T) + (P(X < T) - P(X < t_x))P(Y = 1|X < T) \\ b &= P(X \geq t_x) - (P(X \geq T)P(Y = 1|X \geq T) - (P(X < T) - P(X < t_x))P(Y = 1|X < T)) \\ c &= (P(X < t_x))P(Y = 1|X < T) \\ d &= (P(X < t_x)) - (P(X < t_x))P(Y = 1|X < T) \end{aligned} \tag{2.1}$$

2. $t_x = T$,

$$\begin{aligned} a &= P(X \geq T)P(Y = 1|X \geq T) \\ b &= P(X \geq T) - P(X \geq T)P(Y = 1|X \geq T) \\ c &= P(X < T)P(Y = 1|X < T) \\ d &= P(X < T) - P(X < T)P(Y = 1|X < T) \end{aligned} \tag{2.2}$$

3. $t_x > T$

$$
\begin{aligned}
a &= P(X \geq t_x)P(Y = 1|X \geq T) \\
b &= P(X \geq t_x) - P(X \geq t_x)P(Y = 1|X \geq T) \\
c &= P(X < T)P(Y = 1|X < T) + (P(X < t_x) - P(X < T))P(Y = 1|X \geq T) \\
d &= (P(X < t_x) - (P(X < T)P(Y = 1|X < T) - (P(X < t_x) - P(X < T))P(Y = 1|X \geq T))
\end{aligned}
\tag{2.3}
$$

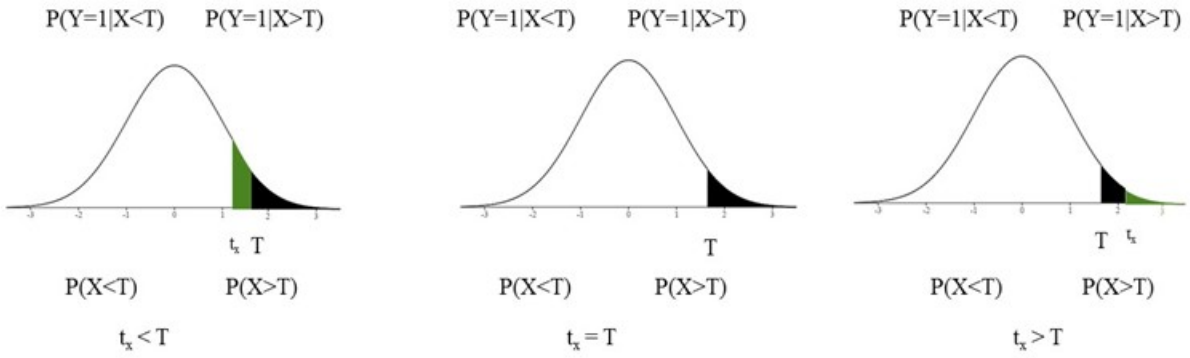

Figure 2.1: Graphical representation of possible thresholds for X presented in equations 1-3.

The numerical values for the statistics in Table 2.2 calculated over the interval $[-4, 4]$ are shown in Figure 2.4 in the Chapter 2 Supplemental Materials section. Methods for which the maximum absolute value for the statistic occurs at the true threshold are considered successful. There are six statistics for which the maximum value occurs at T, namely chi-square, kappa, Youden's, Gini Index, relative risk and odds ratio.

	$Y = 1$	$Y = 0$	
$X \geq T$	$a = P(Y = 1, X \geq T)$ $= P(X \geq T)P(Y = 1 \mid X \geq T)$	$b = P(Y = 0, X > T)$ $= P(X \geq T) - P(X \geq T)P(Y = 1 \mid X \geq T)$	$P(X \geq T)$
$X < T$	$c = P(Y = 1, X < T)$ $= (1 - P(X \geq T))P(Y = 1 \mid X < T)$	$d = P(Y = 0, X < T)$ $= (1 - P(X \geq T)) - (1 - P(X \geq T))P(Y = 1 \mid X < T)$	$P(X < T)$
	$P(Y = 1)$	$1 - P(Y = 1)$	

Table 2.1: Probabilities for a 2x2 contingency table for a binary outcome Y and a continuous variable X thresholded at T

Odds Ratio	Youden's Statistic	Chi-Square
$\frac{ad}{bc}$	$\frac{a}{a+c} + \frac{d}{b+d} - 1$	$\frac{(ad-bc)^2}{(a+b)(c+d)(b+d)(a+c)}$
Kappa Statistic	Relative Risk*	Gini Index
$\frac{(a+d)-((a+b)(a+c)+(c+d)(b+d))}{1-((a+b)(a+c)+(c+d)(b+d))}$	$\frac{a/(a+b)}{c/(c+d)}$	$(P_y(1-P_y)) - (\frac{ab}{a+b} + \frac{cd}{c+d})$
Specificity	Misclassification	Sensitivity
$\frac{d}{b+d}$	$1 - \frac{a+d}{N}$	$\frac{a}{a+c}$
Accuracy Area	Minimax	Minimum ROC**
$\frac{a}{a+c} \cdot \frac{d}{b+d}$	$\max(b, c)$	$\sqrt{(1-\frac{a}{a+c})^2 + (1-\frac{d}{b+d})^2}$

Table 2.2: Formulas for statistics for selecting a threshold for a continuous variable X to discriminate a binary outcome Y based on the probabilities in Table 1
*For cohort study
**Measures the distance from the ROC curve to the point (0,1)

2.2.2 Theoretical confirmation

Based on the numerical evidence presented in section 2.1, the following theorem is conjectured for functions in the first 2 rows of Table 2.2 which include the six statistics for which the maximum occurs at the true threshold.

Theorem 1 *For a continuous random variable X and dichotomous variable Y, given a prevalence of Y ($P(Y = 1)$), and a threshold T such that, $P(Y = 1|X \geq T) > P(Y = 1|X < T)$, the inequality $g(t) < g(T)$ for all $t \neq T$ holds. Here $g(t)$ is any one of the functions shown in the first two rows of Table 2.2. That is, if there exists a true threshold T, the maximum odds ratio, Youden's statistic, chi-square statistic, Gini Index, kappa statistic, or relative risk will occur at T.*

Proof We first consider the case where $P(X > t_x) > P(X \geq T)$.

Consider the case where multiplying both sides of the condition $P(Y = 1|X \geq T) > P(Y = 1|X < T)$ by $P(t_x < X < T)$ yields,

$$P(t_x < X < T)P(Y = 1|X \geq T) > P(t_x < X < T)P(Y = 1|X < T).$$

Replacing $P(t_x < X < T)$ on the LHS with $P(t_x > X) - P(X > T)$ and adding $P(X >$

$T)P(Y = 1|X \geq T)$ to both sides yields,

$$P(t_x > X)P(Y = 1|X \geq T) > P(t_x < X < T)P(Y = 1|X < T) + P(X > T)P(Y = 1|X \geq T).$$

Subtracting $P(X > T)P(Y = 1|X \geq T)^2 + P(t_x < X < T)P(Y = 1|X \geq T)P(Y = 1|X < T)$ from both sides and factoring yields,

$$P(Y = 1|X \geq T)(P(t_x > X) - P(X > T)P(Y = 1|X \geq T) - P(t_x < X < T)P(Y = 1|X < T)) > (1 - P(Y = 1|X \geq T))(P(X > T)P(Y = 1|X \geq T) + P(t_x < X < T)P(Y = 1|X < T)).$$

Dividing both sides by $(P(t_x > X) - P(X > T)P(Y = 1|X \geq T) - P(t_x < X < T)P(Y = 1|X < T))$ and $(1 - P(Y = 1|X \geq T))$ we have,

$$\frac{P(Y = 1|X \geq T)}{(1 - P(Y = 1|X \geq T))} > \frac{(P(X > T)P(Y = 1|X \geq T) + P(t_x < X < T)P(Y = 1|X < T))}{(P(t_x > X) - P(X > T)P(Y = 1|X \geq T) - P(t_x < X < T)P(Y = 1|X < T))}.$$

Multiplying both sides by $(1 - P(Y = 1|X < T))$ yields,

$$\frac{P(Y = 1|X \geq T)(1 - P(Y = 1|X < T))}{(1 - P(Y = 1|X \geq T))} > \frac{(P(X > T)P(Y = 1|X \geq T) + P(t_x < X < T)P(Y = 1|X < T))(1 - P(Y = 1|X < T))}{(P(t_x > X) - P(X > T)P(Y = 1|X \geq T) - P(t_x < X < T)P(Y = 1|X < T))}.$$

Finally, dividing both sides by $P(Y = 1|X < T)$ yields,

$$\frac{P(Y = 1|X \geq T)(1 - P(Y = 1|X < T))}{(1 - P(Y = 1|X \geq T))P(Y = 1|X < T)} > \frac{(P(X > T)P(Y = 1|X \geq T) + P(t_x < X < T)P(Y = 1|X < T))(1 - P(Y = 1|X < T))}{(P(t_x > X) - P(X > T)P(Y = 1|X \geq T) - P(t_x < X < T)P(Y = 1|X < T))P(Y = 1|X < T)}$$

which means $\frac{a_T d_T}{b_T c_T} > \frac{a_{t_x} d_{t_x}}{b_{t_x} c_{t_x}}$ and $OR_{t_x=T} > OR_{t_x>T}$.

Now consider the case where multiplying both sides of the condition $P(Y = 1|X \geq T) > P(Y = 1|X < T)$ by $P(X > t_x) < P(X > T)$ yields the equation,

$$(P(X > t_x) - P(X \geq T)) \cdot P(Y = 1|X \geq T) > (P(X < T) - P(X < t_x))P(Y = 1|X < T)$$

.

Distributing $P(Y = 1|X \geq T)$ on the LHS and adding $P(X \geq T)P(Y = 1|X \geq T)$ to both sides yields,

$$P(X > t_x) \cdot P(Y = 1|X \geq T) > P(X \geq T)P(Y = 1|X \geq T) + (P(X < T) - P(X < t_x))P(Y = 1|X < T).$$

Next, subtracting $P(Y = 1|X \geq T)(P(X \geq T)P(Y = 1|X \geq T) + (P(X < T) - P(X < t_x))P(Y = 1|X < T))$ from both sides and dividing by $(1 - P(Y = 1|X \geq T))$ and $(P(X \geq T)P(Y = 1|X \geq T) + (P(X < T) - P(X < t_x))$ yields,

$$\frac{P(Y = 1|X \geq T)}{(1 - P(Y = 1|X \geq T))} > \frac{(P(X \geq T)P(Y = 1|X \geq T) + (P(X < T) - P(X < t_x))P(Y = 1|X < T))}{(P(X > t_x) - P(X \geq T)P(Y = 1|X \geq T) + (P(X < T) - P(X < t_x))P(Y = 1|X < T))}.$$

Multiplying both sides by $\frac{(1-P(Y=1|X<T))}{P(Y=1|X<T)}$ we have,

$$\frac{P(Y = 1|X \geq T)(1 - P(Y = 1|X < T))}{(1 - P(Y = 1|X \geq T))P(Y = 1|X < T)} > \frac{(P(X \geq T)P(Y = 1|X \geq T) + (P(X < T) - P(X < t_x))P(Y = 1|X < T))(1 - P(Y = 1|X < T))}{(P(X > t_x) - P(X \geq T)P(Y = 1|X \geq T) + (P(X < T) - P(X < t_x))P(Y = 1|X < T))(P(Y = 1|X < T))}.$$

Thus $OR_{t_x=T} > OR_{t_x<T}$. If the expression for $OR_{t_x=T}$ is greater than the expression for $OR_{t_x<T}$ and the expression for $OR_{t_x=T}$ is greater than the expression for $OR_{t_x>T}$ then it shows that the odds ratio is the highest when $t_x = T$.

The proofs of this theorem for the other five statistics are provided in Appendix A.

2.3 Simulations Study

In Section 2, six statistics that are maximized at the true threshold T, were identified. However, if a sample is drawn from a population, it is not clear which of these statistics will most accurately identify the true threshold. To evaluate the ability of the six statistics to recover the true threshold, a simulation study was performed. Sample data sets were generated by first generating a continuous normal random variable $X \sim N(0, 1)$. A binary variable Y was then generated according to the relationship defined in Equation 2.4.

$$P(Y = 1) = P(X \geq T)P(Y = 1|X \geq T) + P(X < T)P(Y = 1|X < T) \tag{2.4}$$

where $P(Y = 1|X > T) > P(Y = 1|X < T)$ and T is the true threshold for dichotomizing X.

Simulations were performed under various scenarios arising from combinations of the parameters: the number of observations in the sample, $N = 250, 500$,or1000, the overall prevalence of Y defined by $P(Y = 1)$, the choice of threshold T for X, strength of association between predictor X and response Y defined by an odds ratio, and case-control or cohort study designs. The probabilities $P(Y = 1|X \geq T)$ and $P(Y = 1|X < T)$ were calculated based on the choice of T, the odds ratio, and the prevalence of Y. For cohort study scenarios, we generated N values of X and Y. For the 1:1 case:control scenarios, we generated 20,000 X and Y values and selected $\frac{N}{2}$ cases and $\frac{N}{2}$ controls. All simulation scenarios are described in Table 2.3. The true parameter values and simulation scenario for each method for $n = 250$ can be found in Supplemental Table S1.

For each simulation scenario and sample size, we generated 500 datasets. The choice of threshold for each method was estimated by calculating the probabilities of a,b,c, and d as described in Table 2.1 for all possible thresholds for X. These probabilities were converted into cell counts by multiplying by the sample size N. We then calculated the associated odds ratio, kappa statistic, chi-square statistic, Youden's statistic, and Gini Index for the 2x2 table corresponding to each unique threshold for X in the observed data. Any threshold for X that resulted in a cell count of less than

Table 2.3: Simulation Scenarios

OR	$P(X \geq T)$	$P(Y = 1)$	$P(Y = 1 \mid X \geq T)$	Scenario
1.5	0.05	0.2	0.268	1
		0.4	0.495	2
	0.2	0.2	0.255	3
		0.4	0.479	4
	0.5	0.2	0.232	5
		0.4	0.448	6
3	0.05	0.2	0.411	7
		0.4	0.654	8
	0.2	0.2	0.363	9
		0.4	0.614	10
	0.5	0.2	0.283	11
		0.4	0.528	12
6	0.05	0.2	0.569	13
		0.4	0.786	14
	0.2	0.2	0.475	15
		0.4	0.735	16
	0.5	0.2	0.326	17
		0.4	0.6	18

1 was eliminated from consideration in order to minimize the influence of extreme values. Across simulation scenarios less than 6% of observations on average were eliminated from consideration in the cohort setting and less than 2% in the case-control setting. The thresholds that corresponded to the maximum value obtained for each statistic were selected as the “best" thresholds. Assessment of how well the maximum of each statistic recovered the true threshold, T, was determined by examining the mean squared error and the bias squared for the estimated threshold across all simulated datasets for all scenarios. All simulations were conducted in R v. 3.2.1 [55].

2.4 Simulation Results

Figure 2.2 shows the results from the simulation study for the case-control study design scenario. Each graph shows the mean squared error (MSE) by bias squared for all statistics described in Table 2.3 for the different combinations of $P(X \geq T)$ and strength of association with Y. The columns in Figure 2.2 show the impact of increasing values for $P(X \geq T)$ and the rows show the impact of increasing strength of association with Y. Three different sample sizes, 250, 500, and

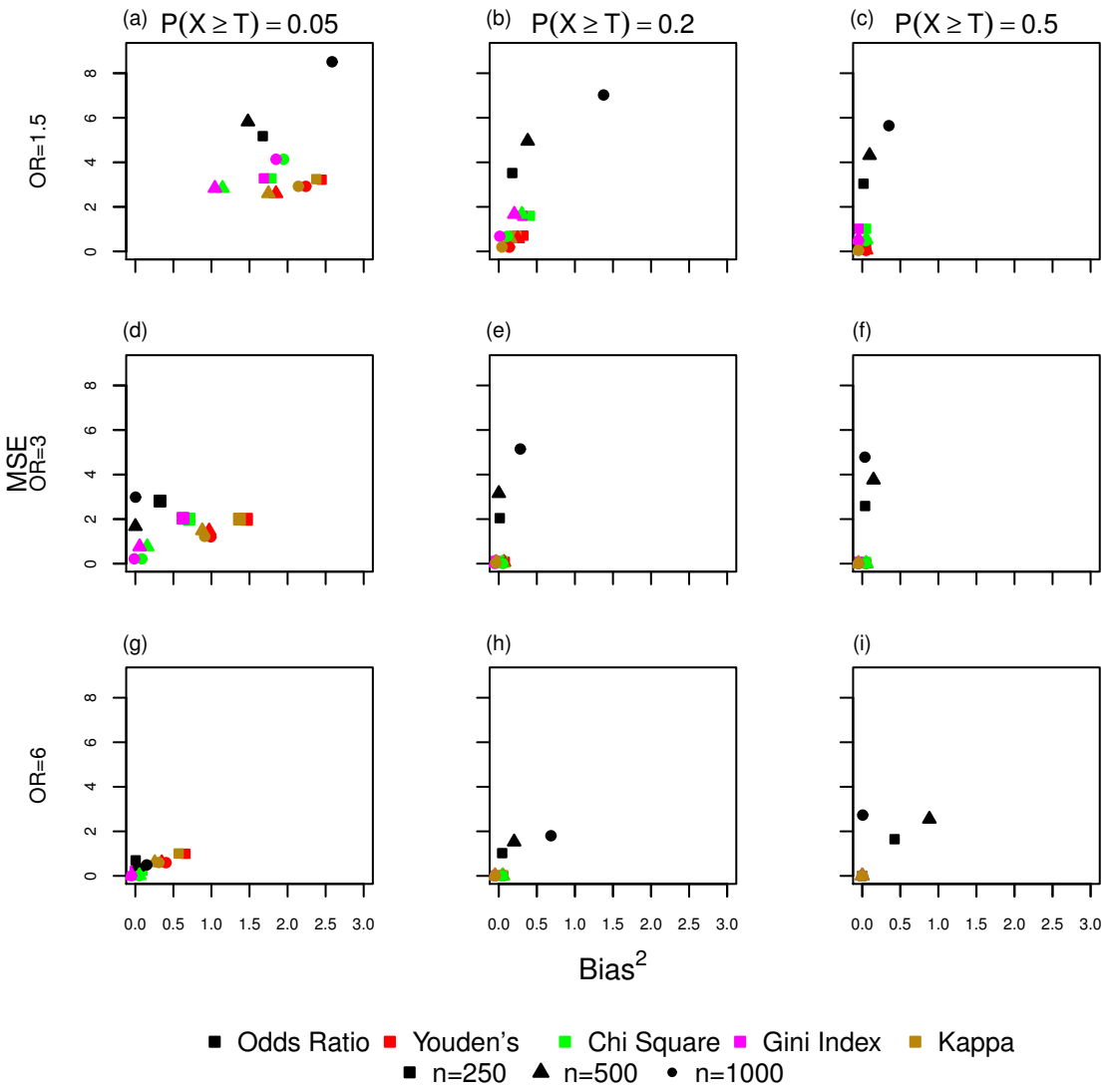

Figure 2.2: Simulation results showing mean-squared error (MSE) by bias^2 under the case-control study design for the estimated threshold obtained by maximizing the statistics: odds ratio, Youden's, chi-square, Gini Index, and kappa. Rows represent strength of association between X and Y and columns represent the probability that the independent variable X is greater than the true threshold T.

1000, are represented by the different shapes, square, triangle, and circle, respectively and each statistic is represented by a different color. As the strength of association between X and Y increases (OR=1.5 to OR=6), all statistics exhibit smaller MSE and bias squared for the estimated threshold indicating that the estimate threshold based on each statistic becomes less variable and biased. As the probability of observing values of X above the true threshold increases, a majority of the methods

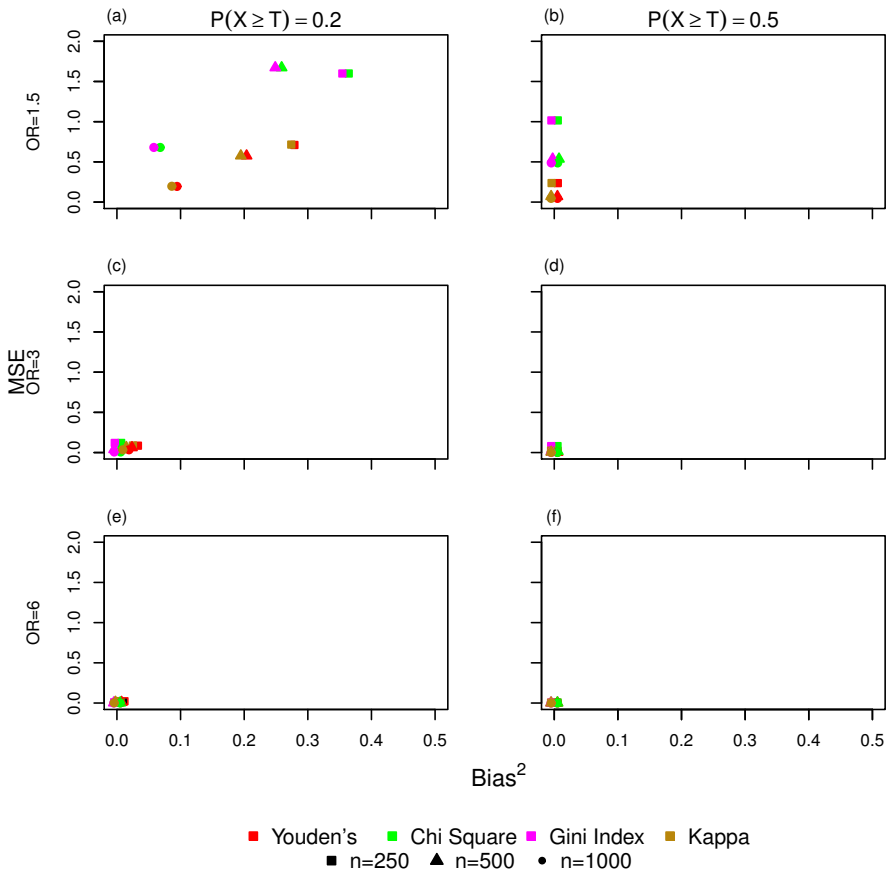

Figure 2.3: Simulation results showing mean-squared error (MSE) by bias^2 under the case-control study design for the estimated threshold obtained by maximizing the statistics: Youden's, chi-square, Gini Index, and kappa, excluding $P(X \geq T)$=0.05. Rows represent strength of association between X and Y and columns represent the probability that the independent variable X is greater than the true threshold T.

show a reduction in bias squared and MSE of the estimated threshold as $P(X \geq T)$ increases from 0.05 to 0.5. The main exception is the odds ratio which does show a decrease in bias for weaker strength of association (OR = 1.5, Figure 2.2 a-c), but which has worse bias as $P(X \geq T)$ increases when the strength of association with Y is large (Figure 2.2 g-i). Additionally, the odds ratio also exhibits an increase in MSE as $P(X \geq T)$ increases with a stronger association between X and Y (Figure 2.2 d-i). As sample size increases, most of the methods show a reduction in MSE and bias^2. The only exception occurs when the strength of association is the weakest and $P(X \geq T)$ is the smallest (Figure 2.2a). Figure 2.3 shows the results for $P(X \geq T)$= 0.2 and $P(X \geq T) = 0.5$

excluding the odds ratio as it is the least effective at recovering the true threshold. When the strength of association is large and $P(X \geq T)$ is > 0.2, the chi-square statistics, Youden's statistic, Gini Index, and kappa statistic all exhibit minimal MSE and bias^2 (Figure 2.3 c-f).

Among the 5 statistics, the odds ratio statistic exhibits the largest MSE and often the largest bias^2 across all simulation scenarios (Figure 2.2 a-i). The bias^2 in the estimated threshold is largest for the odds ratio when the strength of association between X and Y is large (OR=6) and $P(X \geq T) \geq 0.2$ (Figure 2.2h and i). The Gini Index and chi-square statistic perform similarly to one another for all simulation scenarios. In general, both statistics perform well in comparison to the three other statistics when $P(X \geq T)$=0.05 irrespective of the strength of association (OR=1.5, 3, or 6) with the exception of the weakest strength of association between X and Y (OR=1.5) where the kappa and Youden's statistics perform slightly better (Figure 2.2a). The kappa and Youden's statistics also perform similarly to one another across all simulation scenarios and their performance is better than chi-square and Gini Index when $P(X \geq T)$=0.2 or 0.5. The chi-square statistic, Gini Index, Youden's statistic, and kappa statistic all have a squared bias and MSE very near 0 when $P(X \geq T) \geq 0.2$ and the strength of association between X and Y is large (OR> 3).

We also investigated the direction of the bias. Across most of the simulation scenarios, the chi-square statistic, Gini Index, kappa statistic, and Youden's statistic are negatively biased. The only exception is Youden's statistic which has a small positive bias when $P(X \geq T)$ = 0.5 and OR = 6.0. Bias is more variable for the odds ratio. The odds ratio is negatively biased at all sample sizes when the strength of association is small (OR = 1.5). Odds ratio also tends to exhibit negative bias when $P(X \geq T)$=0.5, although this is inconsistent for n = 1000 as strength of association increases. Once P(X>T) is at least 0.2, all methods exhibit negligible bias except for the odds ratio, which has a bias that varies between -0.94 and 0.82.

Simulation scenarios assuming a cohort study design produced very similar results to the case-control scenarios. The relative risk, rather than the odds ratio, was evaluated in the cohort scenarios. Similar to the case-control scenario, the chi-square statistic, Youden's statistics, Gini Index, and kappa statistic all exhibited a reduction in MSE and bias for the estimated threshold as $P(X \geq T)$ increase, strength of association between X and Y increase, and with increasing sample size (Supplemental

Figures 2.5 and 2.6). Similar to what was observed for the odds ratio in the case-control scenario, the relative risk tended to have larger MSE and bias relative to the other four methods with the largest differences observed as strength of association and $P(X \geq T)$ increased. Also similar to the results in the case-control scenario, the chi-square statistic, Gini Index, Youden's statistic, and kappa statistic all have a bias and MSE very near 0 when $P(X \geq T) \geq 0.2$ and the strength of association between X and Y is large (OR$\geq$ 3) (Figure 2.6 c-f). One notable difference between the case-control and cohort scenarios is the performance of the kappa statistic. In the cohort scenarios the kappa statistic selects an estimated threshold with similar or smaller MSE relative to the other four statistics in all scenarios.

2.4.1 Summary of Results

The simulation study examined the performance of the odds ratio, Youden's statistic, chi-square statistic, Gini Index, relative risk and kappa statistic to recover a true threshold, T, for continuous predictor X to discriminate a binary outcome Y. All of these statistics improve on average at finding the true threshold as sample size increases, strength of association between X and Y increases, and as $P(X \geq T)$ increases. The statistic with the most variability in all scenarios is the odds ratio. When the strength of association between X and Y is small, all methods exhibit a larger MSE and bias relative to the truth. When the population odds ratio increases to OR$\geq$ 3, Youden' statistic and kappa statistic exhibit the lowest MSE and bias relative to the odds ratio, chi-square statistic, and Gini Index. Study design had little effect on the performance of the methods.

2.5 Conclusions

Continuous variables are often dichotomized in medical applications to discriminate disease status of a patient population and thereby assist in directing the treatment of a patient. For example, a continuous laboratory value might be dichotomized in order to stratify patients in to disease risk categories in order to make a determination about medication a patient should receive. Additionally, dichotomization of a continuous predictor might be utilized in statistical modeling to simplify

interpretation.

Numerous methods have been described in the literature for dichotomizing a continuous variable to discriminate a binary outcome. In this paper, we provided numerical evidence followed by mathematical proofs that maximizing the odds ratio, relative risk, Youden's statistic, chi-square statistic, Gini Index, and kappa statistic theoretically recover a true threshold for a continuous random variable X, when one exists. In the simulation study, these six statistics exhibited lower MSE and bias as sample size, strength of association, and $P(X \geq T)$ increased. The odds ratio and relative risk statistics were the most variable and exhibited a higher MSE and bias relative to the other methods. If the event is rare (i.e $P(X \geq T)$=0.05), chi-square statistic and Gini Index have the smallest MSE and Bias regardless of strength of association (OR=1.5, 3, or 6). But when $P(X \geq T) > 0.2$, then kappa and Youden's statistic has the smallest MSE and bias. Once there is both a large strength of association (OR=3 or 6) and a high probability for the event ($P(X \geq T)$=0.2 or 0.5), all four are similar. It is our recommendation that odds ratio and relative risk should not be used as they provide the least optimal results and most variable.

We are not discounting the use of other statistics to dichotomize variables. Depending on the situation and type of variable, statistics other than the ones discussed in this paper may be appropriate. Sometimes, it is necessary to use a clinically defined threshold, especially if the focus is on developing a diagnostic test with high sensitivity.

The mixture of binomials for Y defined in Equation 2.4 describes a scenario where there is a steep sigmoidal relationship between a continuous predictor X and dichotomous outcome Y. If the relationship between X and Y is sigmoidal over a large range of X, such as in the case where the probability that Y is 1 follows a logistic relationship with X, the threshold selected by these methods occurs in the most steeply increasing portion of the logistic curve and we would expect greater variability in the selection of a threshold.

2.6 Chapter 2 Supplemental Material

Scenario	OR	Chi	RR	Gini	Youden's	Kappa
1	1.5	0.38	1.36	0.3187	0.021	0.029
2	1.5	0.49	1.25	0.478	0.019	0.023
3	1.5	1.20	1.37	0.318	0.069	0.069
4	1.5	1.63	1.26	0.478	0.066	0.072
5	1.5	1.62	1.38	0.318	0.100	0.064
6	1.5	2.44	1.27	0.478	0.101	0.096
7	3.0	3.67	2.17	0.318	0.066	0.091
8	3.0	3.53	1.69	0.478	0.052	0.062
9	3.0	10.32	2.27	0.318	0.203	0.203
10	3.0	11.92	1.77	0.478	0.178	0.194
11	3.0	10.89	2.43	0.318	0.260	0.167
12	3.0	17.12	1.94	0.478	0.267	0.256
13	6.0	11.21	3.15	0.318	0.115	0.160
14	6.0	8.16	2.07	0.478	0.080	0.094
15	6.0	29.60	3.62	0.318	0.344	0.344
16	6.0	29.24	2.32	0.478	0.279	0.304
17	6.0	24.62	4.37	0.318	0.392	0.251
18	6.0	41.66	3.00	0.478	0.416	0.400

Table 2.4: Specific parameter values for each method given the different scenarios in a case-control setting for sample size of 250.

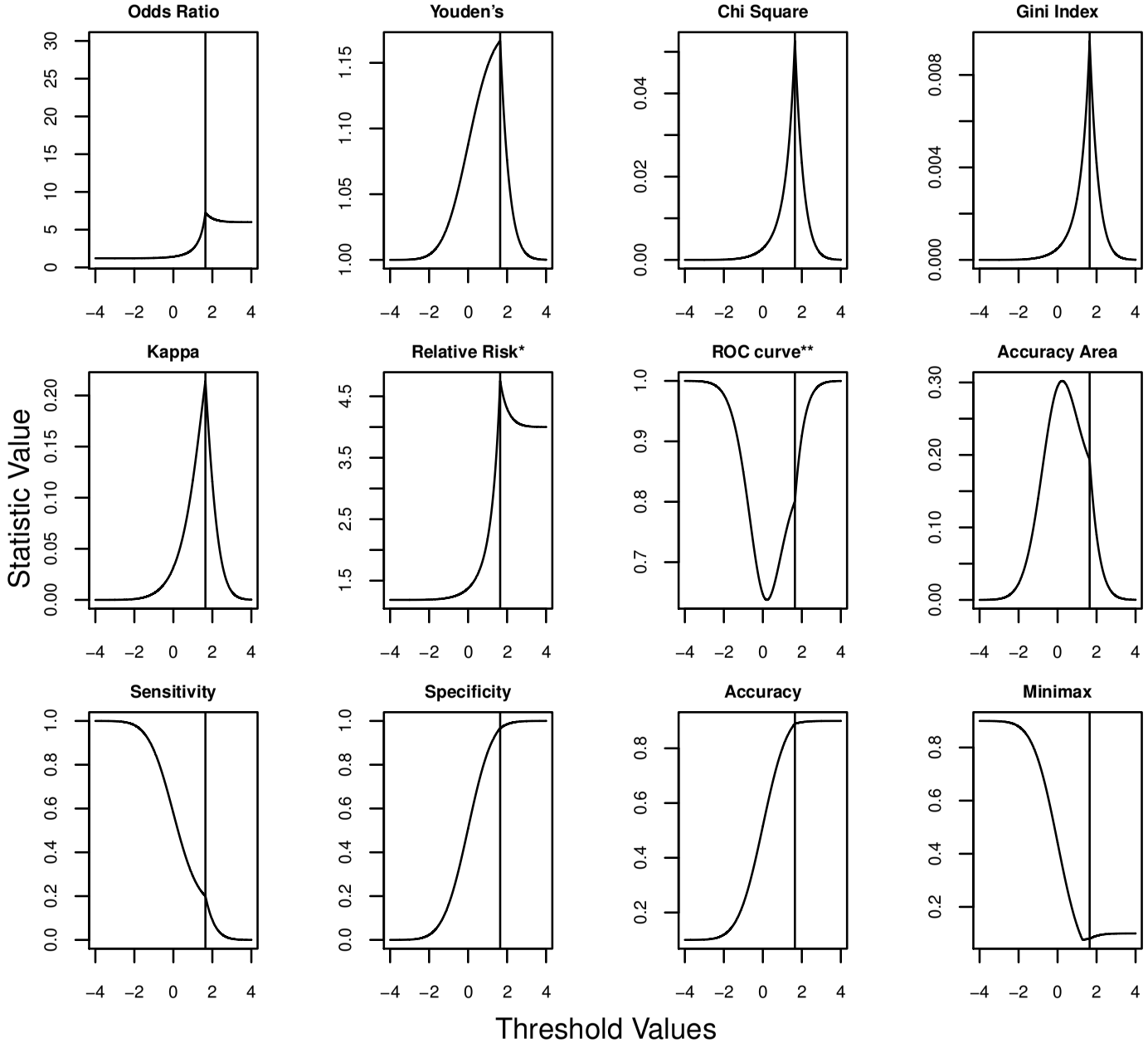

Figure 2.4: Statistic Maximums

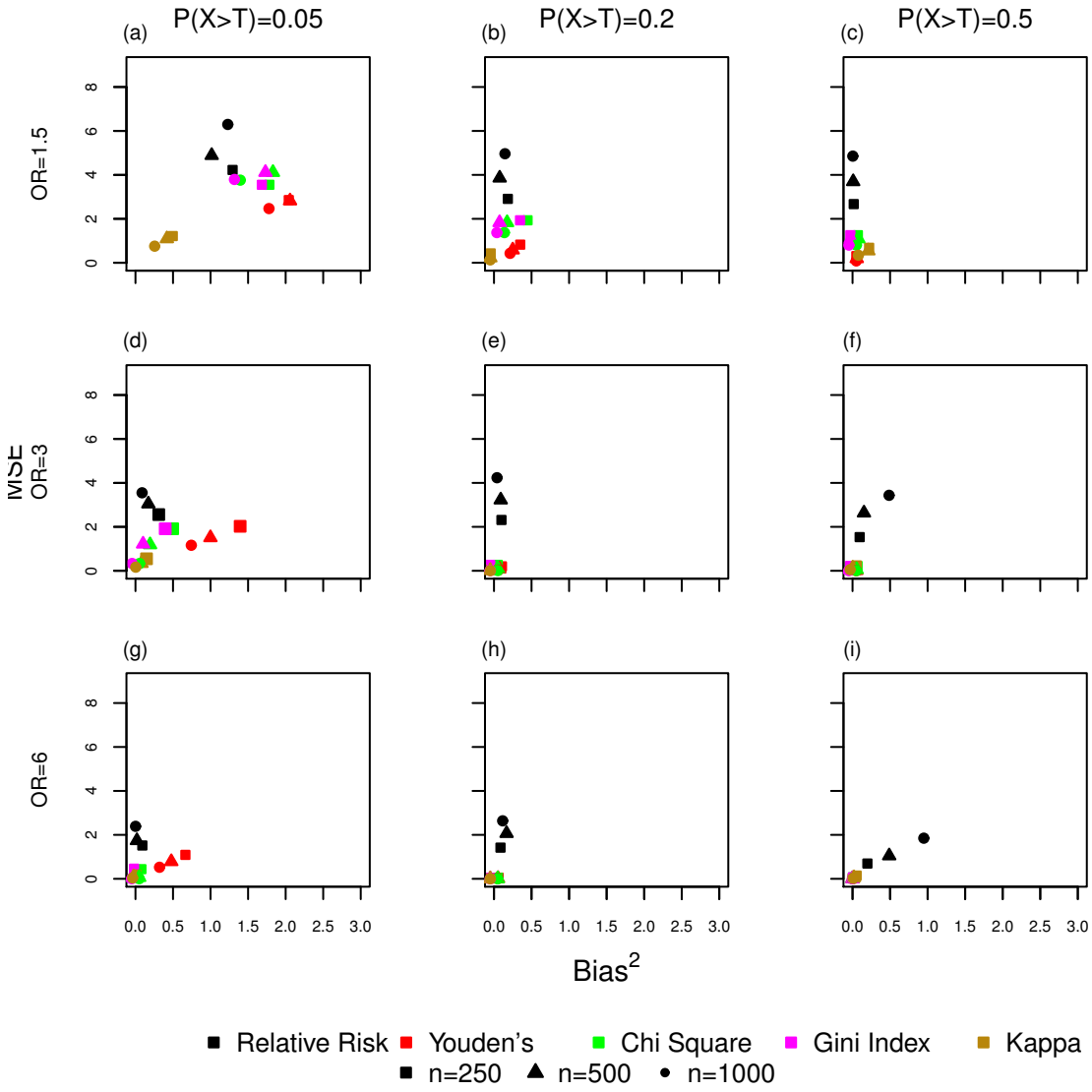

Figure 2.5: Simulation results showing mean-squared error (MSE) by Bias^2 under the cohort study design for the estimated threshold obtained by maximizing the statistics: odds ratio, Youden's, chi-square, Gini Index, and kappa. Rows represent strength of association between X and Y and columns represent the probability that the independent variable X is greater than the true threshold T.

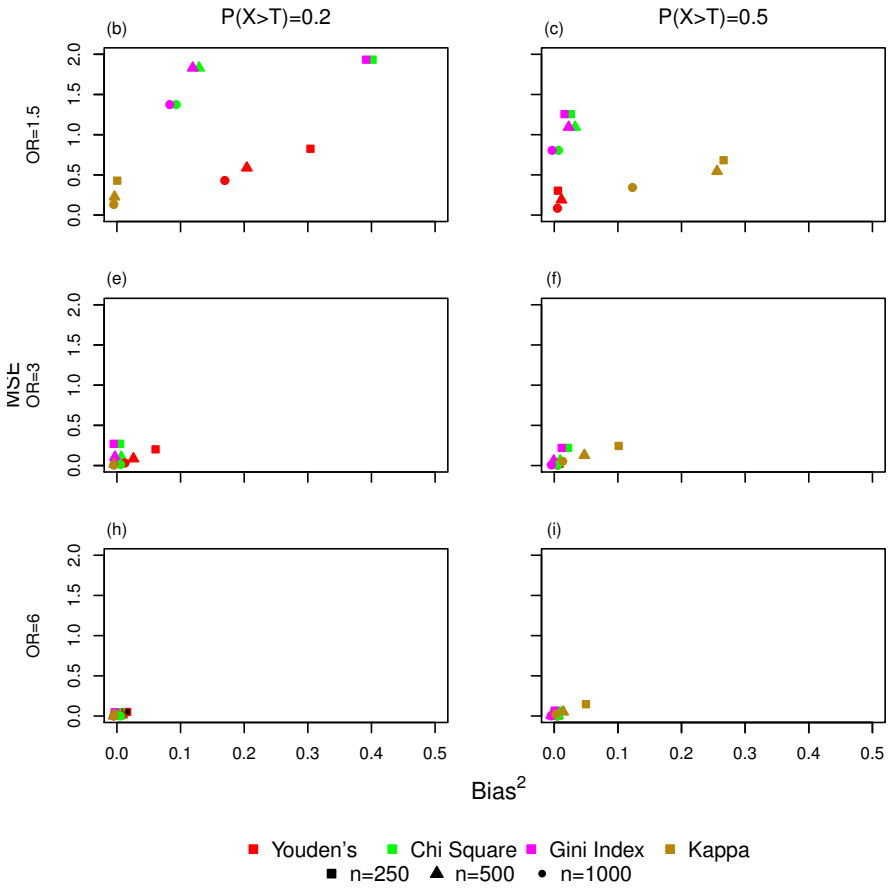

Figure 2.6: Simulation results showing mean-squared error (MSE) by Bias^2 under the cohort study design for the estimated threshold obtained by maximizing the statistics: Youden's, chi-square, Gini Index, and kappa, excluding $P(X \geq T)$=0.05. Rows represent strength of association between X and Y and columns represent the probability that the independent variable X is greater than the true threshold T.

A Comparison of Joint Dichotomization and Single Dichotomization of Interacting Variables

3.1 Introduction

In medicine, continuous variables are often dichotomized to develop diagnostic and prognostic tools to aid in patient care. For example, patients with a total cholesterol level greater than 200mg/dL are believed to have increased risk for cardiovascular disease (CVD) and, therefore, may be prescribed a cholesterol lowering drug [4, 31]. A growing body of evidence suggests that complex diseases such as CVD, may be influenced by the interactions between multiple genetic, clinical, and environmental factors [1, 11, 12]. If disease risk, progression, or response to therapy are influenced by the interaction of two or more factors rather than by each factor independently, then dichotomizing these factors separately may result in less than optimal choices of threshold for both factors. Also, if continuous factors are interacting with other variables yet are dichotomized separately, their interaction with each other and with disease outcome may never be identified. For example, in studies examining gene association with disease outcome researchers test the association between disease status with each gene individually and genes without a strong association are eliminated from further study [56]. It is possible, however, that one or more genes are associated with disease only in the presence of another gene or environmental factor. Thus, considering multiple factors simultaneously may lead to identification of genetic and environmental factors associated with disease outcome that might otherwise be missed. If the factors of interest are continuous and must be dichotomized for clinical or statistical reasons,they should also be dichotomized simultaneously (jointly) in order to preserve their association with each other and the outcome.

There are many methods for finding an optimal threshold to dichotomize a single continuous variable for discriminating a binary outcome. However, there is limited methodology described in the literature to simultaneously optimize the thresholds for two or more variables to discriminate a binary outcome. For example, decision tree methods such as Classification and Regression Trees

(CART) have the ability to identify thresholds ("cut-points") for more than one continuous variable but these dichotomization processes are done sequentially rather than simultaneously.

In this chapter, we describe an interaction in which only the presence of two or more variables lead to increased risk of disease and not any single variable alone. We also describe an algorithm for jointly dichotomizing those variables to discriminate a binary outcome. Section 3.2 of this chapter describes the framework for an interaction term and gives numerical justification for joint dichotomization. In Section 3.3, we will provide theoretical proof that maximizing the statistics identified in chapter 2 finds the true threshold. Section 3.4 describes the algorithm for joint thresholding. Section 3.5 presents the results of a simulation study designed to evaluate the impact of the location of the true thresholds, sample size, and strength of association between the binary outcome and the interaction on the ability of the methods described in chapter 2 to correctly estimate the threshold. The simulation study shows that there is less variability and bias in the selection of thresholds when they are chosen jointly rather than individually for the statistics we identified in chapter 2. In section 3.6, we will discuss the implications of the simulation results.

3.2 Case for Joint Dichotomization

This section provides an empirical and theoretical comparison of six methods for selecting thresholds to dichotomize two or more continuous variables, $X = (X_1, X_2, \ldots, X_p)$, to discriminate a binary outcome, Y, by jointly or singly selecting the thresholds, $T = (t_{x_1}, t_{x_2}, \ldots, t_{x_p})$, for each variable when the interaction between the variables is associated with the outcome. The threshold for a continuous variable or set of variables can be selected by maximizing or minimizing specific statistics, which can estimated from a 2x2 contingency table for the binary outcome Y and dichotomized X.

In the previous chapter, we showed that when a true threshold for a continuous variable exists that discriminates a binary outcome, dichotomization based on maximizing one of six statistics- odds ratio, relative risk, Youden's statistic, chi-square statistic, Gini Index and kappa statistic- theoretically correctly recovers the true threshold given the relationship between Y and a continuous variable X

has the relationship defined in Equation 2.4:

$$P_{Y=1} = P_{X \geq T} P_{Y=1|X \geq T} + P_{X<T} P_{Y=1|X<T}$$

where $P_{Y=1|X>T} > P_{Y=1|X<T}$ and T is the true threshold for dichotomizing X.

For this chapter, we extend this definition to include two variables, X_1 and X_2. We describe the interaction between them as:

$$P_{Y=1} = P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} + (P_{X_1 \geq T_1, X_2 < T_2} + P_{X_1 < T_1, X_2 \geq T_2} + P_{X_1 < T_1, X_2 < T_2}) P_{Y|F} \quad (3.1)$$

where $P_{Y|T} = P_{Y=1|X_1 \geq T_1, X_2 \geq T_2}$ and $P_{Y|F} = P_{Y=1|X_1 \geq T_1, X_2 < T_2 \vee X_1 < T_1, X_2 \geq T_2 \vee X_1 < T_1, X_2 < T_2}$ and $P_{Y|T} > P_{Y|F}$ Thus when $P_{Y=1, X_1 \geq T_1, X_2 \geq T_2}$ we have $P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y=1|X_1 \geq T_1, X_2 \geq T_2}$ and when $P_{Y=0, X_1 \geq T_1, X_2 \geq T_2}$ we have $P_{X_1 \geq T_1, X_2 \geq}(1 - P_{Y=1|X_1 \geq T_1, X_2 \geq T_2})$. The variable X_1 is dichotomized to 0 if $X_1 < T_2$ or if $X_2 < T_2$. This means the probability of X_1 dichotomized to 0 and $P_{Y=1}$ is the sum of three joint probabilities, $P_{Y=1|X_1 < T_1, X_2 \geq T_2} + P_{Y+1|X_1 < T_1, X_2 < T_2} + P_{Y=1|X_1 \geq T_1, X_2 < T_2}$ which can also be written as $(P_{X_1 < T_1, X_2 \geq T_2} + P_{X_1 < T_1, X_2 < T_2} + P_{X_1 \geq T_1, X_2 < T_2}) P_{Y=1|X_1 < T_1 \vee X_2 < T_2}$. Finally, when X_1 is dichotomized to 0 and $P_{Y=0}$ we again have the sum of three joint probabilities, $(P_{X_1 < T_1, X_2 \geq T_2} + P_{X_1 \geq T_1, X_2 < T_2} + P_{X_1 < T_1, X_2 < T_2}) - (P_{X_1 < T_1, X_2 \geq T_2} + P_{X_1 \geq T_1, X_2 < T_2} + P_{X_1 < T_1, X_2 < T_2}) P_{Y=1|X_1 < T_1 \vee X_2 < T_2}$ These probabilities are summarized in Table 3.2

If Y is associated with $\mathbf{X}$ through an interaction, then $P_{Y=1}$ is larger when in the presence of the interaction. For this paper, an interaction between two or more variables means that there is an increased risk of $P_{Y=1}$ when both or all variables are present.

3.2.1 Numeric Investigation of Single and Joint Thresholding

This section provides an empirical examination of the ability of joint and single thresholding to correctly identify a true thresholds, $\mathbf{T}$, in the case where two predictors X_1 and X_2 are associated with a binary outcome Y through the relationship defined in Equation 3.1. We say a variable X_1 is dichotomized singly when the threshold for X, t_{x_1}, is selected by choosing the value of t_{x_1} that

	$Y=1$	$Y=0$	
$X_1 \geq T_1$	$\begin{aligned} a_S &= P(Y=1, X_1 \geq T_1) \\ &= P_{X_1 \geq T_1} P_{Y \mid X_1 \geq T_1} \\ &= P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y \mid T} + P_{X_1 \geq T_1, X_2 < T_2} P_{Y \mid F} \end{aligned}$	$\begin{aligned} b_S &= P(Y=0, X_1 \geq T_1) \\ &= P_{X_1 \geq T_1}(1 - P_{X_2 \geq T_2} P_{Y \mid T} - P_{X_2 < T_2} P_{Y \mid F}) \end{aligned}$	$P_{X_1 \geq T_1}$
$X_1 < T_1$	$\begin{aligned} c_S &= P(Y=1, X_1 < T_1) \\ &= P_{X_1 < T_1} P_{Y \mid F} \end{aligned}$	$\begin{aligned} d_S &= P(Y=0, X_1 < T_1) \\ &= P_{X_1 < T_1}(1 - P_{Y=1 \mid F}) \end{aligned}$	$P_{X_1 < T_1}$
	$P_{Y=1}$	$P_{Y=0}$	

Table 3.1: The 2x2 contingency table for continuous variable X_1 and dichotomous outcome Y where X_1 is singly thresholded at T_1 when the relationship between X_1, X_2, and Y is described in Equation 3.1

	$Y=1$	$Y=0$	
$X_1 \geq T_1, X_2 \geq T_2$	$\begin{aligned} a_J &= P(Y=1, X_1 \geq T_1, X_2 \geq T_2) \\ &= P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y \mid X_1 \geq T_1, X_2 \geq T_2} \end{aligned}$	$\begin{aligned} b_J &= P(Y=0, X_1 \geq T_1, X_2 \geq T_2) \\ &= P_{X_1 \geq T_1, X_2 \geq T_2} \\ &- P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y \mid X_1 \geq T_1, X_2 \geq T_2} \end{aligned}$	$P_{X_1 \geq T_1, X_2 \geq T_2}$
$X_1 < T_1 \vee X_2 < T_2$	$\begin{aligned} c_J &= P(Y=1, X_1 < T_1, X_2 \geq T_2) \\ &+ P(Y=1, X_1 \geq T_1, X_2 < T_2) \\ &+ P(Y=1, X_1 < T_1, X_2 < T_2) \\ &= (P_{X_1 < T_1, X_2 \geq T_2} \\ &+ P_{X_1 \geq T_1, X_2 < T_2} \\ &+ P_{X_1 < T_1, X_2 < T_2}) P_{Y \mid X_1 < T_1 \vee X_2 < T_2} \end{aligned}$	$\begin{aligned} d_J &= P(Y=0, X_1 < T_1, X_2 \geq T_2) \\ &+ P(Y=0, X_1 \geq T_1, X_2 < T_2) \\ &+ P(Y=0, X_1 < T_1, X_2 < T_2) \\ &= (P_{X_1 < T_1, X_2 \geq T_2} \\ &+ P_{X_1 \geq T_1, X_2 < T_2} \\ &+ P_{X_1 < T_1, X_2 < T_2}) \\ &- (P_{X_1 < T_1, X_2 \geq T_2} \\ &+ P_{X_1 \geq T_1, X_2 < T_2} \\ &+ P_{X_1 < T_1, X_2 < T_2}) P_{Y=1 \mid X_1 < T_1 \vee X_2 < T_2} \end{aligned}$	$\begin{aligned} &P_{X_1 > T_1, X_2 < T_2} \\ &+ P_{X_1 < T_1, X_2 \geq T_2} \\ &+ P_{X_1 < T_1, X_2 < T_2} \end{aligned}$
	$P_{Y=1}$	$P_{Y=0}$	

Table 3.2: The 2x2 contingency table for continuous variables X_1 and X_2 and dichotomous outcome Y where X_1 and X_2 are jointly thresholded at T_1 and T_2 respectively

Odds Ratio	Youden's Statistic	Chi-Square
$\frac{ad}{bc}$	$\frac{a}{a+c}+\frac{d}{b+d}-1$	$\frac{(ad-bc)^2}{(a+b)(c+d)(b+d)(a+c)}$
Kappa Statistic	Relative Risk*	Gini Index
$\frac{(a+d)-((a+b)(a+c)+(c+d)(b+d))}{1-((a+b)(a+c)+(c+d)(b+d))}$	$\frac{a/(a+b)}{c/(c+d)}$	$(P_y(1-P_y))-(\frac{ab}{a+b}+\frac{cd}{c+d})$

Table 3.3: Formulas for statistics for selecting a threshold for a continuous variable X to discriminate a binary outcome Y based on the probabilities in Table 3.1 *For cohort study designs only

maximizes one of the six statistics in Table 3.3 using the observed values of Y and X_1 ignoring X_2. Joint dichotomization is defined as selecting the thresholds, t_{x_1} and t_{x_2}, for X_1 and X_2 such that one of the selected statistics in Table 3.3 is maximized based on a_J, b_J, c_J and d_J defined in Table 3.2 estimated from observed X_1, X_2 and Y. We select the threshold for X_1 while setting the threshold for X_2 at the truth, $t_{x_2} = T_2$.

To show this numerically, we consider the case where $\mathbf{X} \sim N_2(\mathbf{0}, I_2)$ and $P_{X_1} > T_1 = 0.3, P_{X_2} > T_2 = 0.2, P_{Y=1} = 0.1, P_{Y|T} = 0.2$, and $P_{Y|F} = 0.094$. We then calculate each statistic in Table 3.3 for joint thresholding the probabilities defined in Table 3.2 at each combination of values of X_1 and X_2 in the interval $[-4, 4]$ in increments of 0.001 including the true thresholds, T_1 and T_2.

For single thresholding, we find the threshold of X_1 without considering the value of X_2. In order to calculate the six statistics for different possible thresholds, t_{x_1}, we consider the three cases when $t_{x_1} < T$,$t_{x_1} = T$,and $t_{x_1} > T$. The cell probabilities for singly thresholding X_1 for these three case are shown below.

1. $t_{x_1} = T_1$

$$
\begin{aligned}
a &= P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} + P_{X_1 \geq T_1, X_2 < T_2} P_{Y|F} \\
b &= P_{X_1 \geq T_1} - (P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|F} + P_{X_1 \geq T_1, X_2 < T_2} P_{Y|F}) \\
c &= P_{X_1 < T_1}(P_{Y|F}) \\
d &= P_{X_1 < T_1}(1 - (P_{Y|F}))
\end{aligned}
\tag{3.2}
$$

2. $t_{x_1} < T_1$

$$
\begin{aligned}
a &= P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} + (P_{X_1 \geq T_1} P_{X_2 < T_2} + (P_{X_1 \geq t_{X_1}} - P_{X_1 > T_1})) P_{Y|F} \\
b &= P_{X_1 \geq t_{X_1}} - (P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} + (P_{X_1 \geq T_1} P_{X_2 < T_2} + (P_{X_1 \geq t_{X_1}} - P_{X_1 > T_1})) P_{Y|F}) \\
c &= (P_{X_1 < t_{X_1}, X_2 < T_2} + P_{X_1 < t_{X_1}, X_2 > T_2}) P_{Y|F} \\
d &= (P_{X_1 < t_{X_1}, X_2 < T_2} + P_{X_1 < t_{X_1}, X_2 > T_2})(1 - P_{Y|F})
\end{aligned}
\tag{3.3}
$$

3. $t_{x_1} > T_1$:

$$
\begin{aligned}
a &= P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} + P_{X_1 \geq T_1, X_2 < T_2} P_{Y|F} \\
b &= P_{X_1 \geq t_{X_1}} - (P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} + P_{X_1 \geq T_1, X_2 < T_2} P_{Y|F}) \\
c &= P_{Y=1} - (P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} + P_{X_1 \geq T_1, X_2 < T_2} P_{Y|F}) \\
d &= (1 - P_{Y=1}) - (P_{X_1 \geq t_{X_1}} - (P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} + P_{X_1 \geq T_1, X_2 < T_2} P_{Y|F}))
\end{aligned}
\tag{3.4}
$$

Similar to joint thresholding, for single thresholding calculate the six statistics in Table 3.3 over the range of thresholds for X_1 in the interval $[-4, 4]$ in increments of .001.

Figure 3.1 shows the value of the six statistics for every value of t_{x_1} considered in the interval [-4,4]. The dashed line represents single thresholding and the solid line represents values of the six statistics for different values of t_{x_1} when $t_{x_2} = T_2$. These plots confirm that the true threshold, T_1, for X_1 occurs at the absolute maximums for these statistics when thresholding singly or jointly. Additionally, Figure 3.1 shows that the maximum value for each statistic is smaller when singly thresholding X_1 independent of X_2 versus jointly thresholding X_1 and X_2 when the association between Y with X_1 and X_2 follows the relationship defined in Equation 3.1

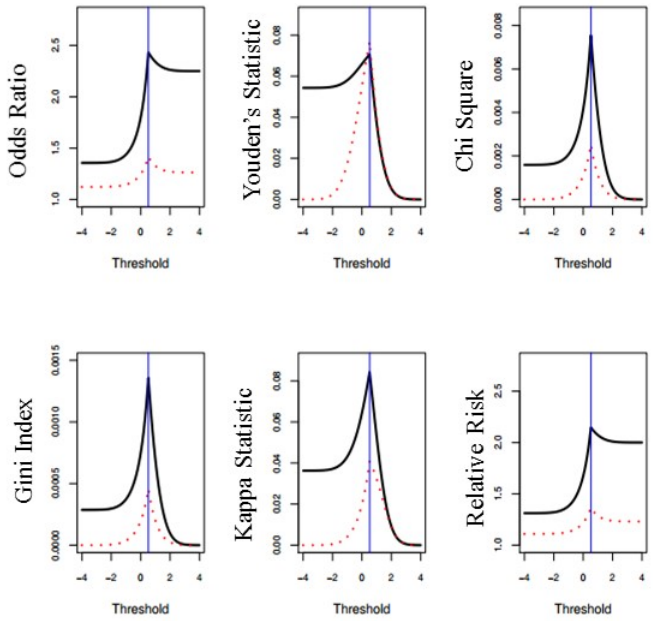

Figure 3.1: Values of statistics from Table 2.2.1 for different thresholds, t_{x_1}; for X_1 under single or joint thresholding in the case where two continuous variables X_1 and X_2 are associated with a binary outcome Y the relationship in the Equation 3.1. Here $X \sim N_2(0, I_2)$,$P_{X_1 \geq T_1} > 0.2$,$P_{X_2 \geq T_2} > 0.2$, $P_{Y=1} = 0.2$, $P_{Y|T} =$,and $P_{Y|F} =$. The solid line represents the value of each statistic for values of t_{x_1} in $[-4, 4]$ under joint thresholding where $t_{x_2} = T_2$. The dashed line represents the values of each statistic for values of t_{x_1} under single thresholding. The vertical line occurs at the true threshold T_1 for X_1.

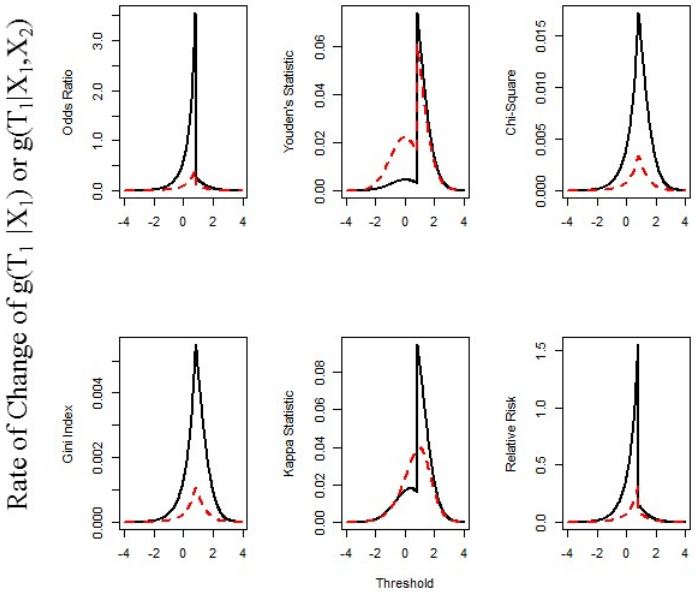

Figure 3.2: Numeric estimation of the first derivative of the six statistic from Table 3.3 for different values of threshold,t_{x_1} for continuous variable X_1 under single or joint dichotomization. Here $P_{X_1 \geq T_1} = 0.2$, $P_{X_2 \geq T_2} = 0.2$,$P_{Y=1} = 0.2$ and $OR = 3$. Under joint dichotomization we assume that t_{x_1} varies while $t_{x_2} = T_2$

3.3 Theoretical confirmation

Define $g(T_1|X_1)$ and $g(T_2|X_2)$as the functions for odds ratio, relative risk, chi-square statistic, gini index, Youden's statistic, and kappa statistic under single thresholding of X_1 and X_2 respectively as defined in Table 3.1. Also define $g(T_k|X_1, X_2)$ as the function for odds ratio, relative risk, chi-square statistic, Gini Index, Youden's statistic, and kappa statistic from Table 3.3 for the k^{th} threshold $(k = 1, 2)$ under joint thresholding of X_1 and X_2 as defined in Table 3.2. Additionally, we say that the relationship between two continuous variables X_1 and X_2 and a dichotomous outcome Y is defined as an interaction of additional risk of Y and occurs only when both conditions $X_1 \geq T_1$ and $X_2 \geq T_2$ are met. In this case, the association between outcome Y and continuous variables X_1 and X_2 is defined by Equation 3.1. Based on the numeric evaluation in Section 2, the following theorems are conjectured.

Theorem 3.1 *For continuous variables X_1 and X_2 and a dichotomous variable Y with prevalence $P_{Y=1}$ and thresholds T_1 and T_2 such that $P_{Y|T} > P_{Y|F}$ (Equation 3.1), then the inequality $g(t_1|X_1) < g(T_1|X_1)$ for all $t_1 \neq T_1$ holds where $g(T|X_1)$ is any of the six statistics defined in Table 3.3.*

Proof: For this case, $g(T|X_1)$ is odds ratio. The statement of the theorem is equivalent to saying $\frac{a_t d_t}{b_t c_t} < \frac{a_T d_T}{b_T c_T}$ where a_T, b_T, c_T and d_T represent the cell values of a 2x2 contingency table determined by thresholding at the true threshold,T, and a_t, b_t, c_t and d_t are cell values determined by a threshold t such that $t \neq T$. First we prove the theorem for $t_{x_1} < T$ and begin with the given statement

$$P_{Y|T} > P_{Y|F}$$

Multiply both sides by $P_{t_{x_1}<X_1<T_1, X_2 \geq T_2}$

$$P_{t_{x_1}<X_1<T_1, X_2 \geq T_2} P_{Y|T} > P_{t_{x_1}<X_1<T_1, X_2 \geq T_2} P_{Y|F}$$

Multiply both sides by $P_{X_1 \geq T_1, X_2 \geq T_2} = (P_{X_1 \geq T_1} - P_{X_1 \geq T_1, X_2 < T_2})$ and note $P_{t_{x_1}<X_1<T_1, X_2 \geq T_2} = -(P_{X_1 \geq T_1} - P_{X_1 \geq t_{x_1}})$

$$P_{X_1\geq T_1,X_2\geq T_2}P_{Y|T}(P_{X_1\geq T_{X_1}}-P_{X_1\geq T_1})>P_{Y|F}(P_{X_1\geq T_1}-P_{X_1\geq T_1,X_2<T_2})P_{t_{x_1}<X_1<T_1,X_2\geq T_2}$$

Distribute, add $P_{X_1\geq T_1,X_2\geq T_2}P_{Y|T}P_{X_1\geq T_1}$ and $P_{Y|F}P_{X_1\geq T_1,X_2<T_2}P_{X_1\geq t_{x_1}}$

$$P_{X_1\geq t_{x_1}}(P_{X_1\geq T_1,X_2\geq T_2}P_{Y|T}+P_{X_1\geq T_1,X_2<T_2}P_{Y|F})>$$
$$P_{X_1\geq T_1}P_{X_1\geq T_1,X_2\geq T_2}P_{Y|T}+P_{X_1\geq T_1}(P_{X_1\geq T_1,X_2<T_2}+P_{t_{x_1}<X_1<T_1,X_2\geq T_2})P_{Y|F}$$

Dividing and comparing terms results in $\frac{a_T d_T}{b_T c_T}>\frac{a_t d_t}{b_t c_t}$ for $t<T$. A similar proof follows for $t>T$

Theorem 3.2 *For continuous variables X_1 and X_2 and a dichotomous variable Y with prevalence $P_{Y=1}$ and thresholds T_1 and T_2 such that $P_{Y|T}>P_{Y|F}$ (Equation 3.1), the rate of convergence to T_1 is faster when jointly thresholding compared to single thresholding. That is,*

$$\frac{\partial g(T_i|X_1,X_2)}{\partial T_i}>\frac{\partial g(T_i|X_i)}{\partial T_i}$$

for either $i=1$or 2 when g is one of the six statistics defined in Table 3.3.

To prove this theorem, we first prove the following lemma.

Lemma 3.1 *For continuous variables X_1 and X_2 and a dichotomous variable Y with prevalence $P_{Y=1}$, and thresholds T_1 and T_2 if $P_{Y|T}>P_{Y|F}$ then for functions g defined earlier, $g(T|X_1,X_2)>g(T|X_1)$ where $g(T|X_1,X_2)$ is defined using the joint events (a_J,b_J,c_J,d_J) and $g(T|X_1)$ uses the marginal events (a_S,b_S,c_S,d_S). We conjecture that this Lemma will extend to the case of p continuous variables where the p variables are associated with dichotomous outcome Y through their interaction. This proof can be shown through induction.*

Proof:

For the case where $g(t)$ is the odds ratio, the statement of the lemma is equivalent to the claim that

$$\frac{a_J d_J}{b_J c_J} > \frac{a_S d_S}{b_S c_S}$$

where (a_S, b_S, c_S, d_S) are defined by the probabilities defined in Table 3.1 for the single thresholding case and (a_J, b_J, c_J, d_J) are the cell probabilities for the joint thresholding case defined in Table 3.2 for the given thresholds T_1 and T_2. To prove the lemma, consider the inequality $P_{Y|T} > P_{Y|F}$ and multiply both sides by $P_{X_1 \geq T_1, X_2 < T_2}$ and $P_{X_1 \geq T_1, X_2 < T_2} = P_{X_1 \geq T_1} - P_{X_1 \geq T_1, X_2 \geq T_2}$ which yields

$$P_{Y|T}(P_{X_1 \geq T_1} - P_{X_1 \geq T_1, X_2 \geq T_2}) > P_{Y|F} P_{X_1 \geq T_1, X_2 < T_2}$$

Now adding $P_{Y|T} P_{X_1 \geq T_1, X_2 \geq T_2} - P_{Y|T}(P_{Y|T} P_{X_1 \geq T_1, X_2 \geq T_2} + P_{Y|F} P_{X_1 \geq T_1, X_2 < T_2})$ to both sides and factoring and simplifying yields

$$\frac{P_{Y|T}}{(1 - P_{Y|T})} > \frac{P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} + P_{X_1 \geq T_1, X_2 < T_2} P_{Y|F}}{(P_{X_1 \geq T_1} - P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} + P_{X_1 \geq T_1, X_2 < T_2} P_{Y|F})}$$

Multiply both sides by $\frac{(1 - P_{Y|F})}{P_{Y|F}}$

$$\frac{P_{Y|T}(1 - P_{Y|F})}{(1 - P_{Y|T}) P_{Y|F}} > \frac{P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} + P_{X_1 \geq T_1, X_2 < T_2} P_{Y|F}(1 - P_{Y|F})}{(P_{X_1 \geq T_1} - P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} + P_{X_1 \geq T_1, X_2 < T_2} P_{Y|F}) P_{Y|F}}$$

Rearranging the terms and recognizing the factors yields

$$\frac{a_J d_J}{b_J c_J} > \frac{a_S d_S}{b_S c_S}$$

Thus,

$$OR_J > OR_S$$

Theorem 3.1 and Lemma 3.1 are also confirmed by the numeric findings shown in Figure 3.1. The proofs for Theorem 3.1 and Lemma 3.1 for some of the other methods can be found in Appendix B. Lemma 3.1 demonstrated that $g(T_i|X_1, X_2) > g(T_i|X_i)$ for $i = 1$ or 2 and for all

T_1, T_2 fixing $t_{X_i} = T_i$. Therefore, the proof of Theorem 3.2 follows. We demonstrate Theorem 3.2 further using a numeric approach shown in Figure 3.2. In Figure 3.2, we examine the rate of change in $g(t_{x_1}|X_1)$ and $g(t_{x_1}|X_1, X_2)$ at $t_{x_2} = T_2$ for small changes in t_{X_1}. The rate of change is calculated for single thresholding as $g(t_{X_1} + 0.001|X_1) - g(t_{X_1}|X_1)$ and for joint thresholding as $g(t_{X_1} + 0.001|X_1, X_2) - g(t_{X_1}|X_1, X_2)$ for t_{X_1} over the range $[-4, 4]$. The solid line is the rate of change under joint thresholding and the dashed line is the rate of change under single thresholding. For all six statistics in Table 3.3, the rate of change near T_1 is faster for joint thresholding relative to single thresholding.

3.4 Joint thresholding algorithm

Box 1: Algorithm for singly thresholding X_1

1. Order the values of continuous variable X to yield oX

2. For each value of X in sample size n, calculate the cell counts for a 2x2 contingency table as follows:

$$\begin{aligned} a_i &= \sum_{i=1}^{n} I(X_i \geq oX_i) \wedge I(Y = 1) \\ b_i &= \sum_{i=1}^{n} I(X_i \geq oX_i) \wedge I(Y = 0) \\ c_i &= \sum_{i=1}^{n} I(X_i < oX_i) \wedge I(Y = 1) \\ d_i &= \sum_{i=1}^{n} I(X_i < oX_i)) \wedge I(Y = 0) \end{aligned} \tag{3.5}$$

3. Missing cells are imputed with a value of 0.5 in order to mitigate extreme values and avoid undefined calculations.

4. Select the value oX_i that maximizes the statistic $g(t|X_1)$, where $g(t)$ is one of the six

statistics in Table 3.1. For example,

$$OR_i = \frac{a_i d_i}{b_i c_i}$$

Now we propose an algorithm to jointly identify the best combination of thresholds t_{x_1} and t_{x_2} for X_1 and X_2 to discriminate a binary outcome Y. The proposed algorithm is shown in Box 2.

Box 2: Algorithm for jointly thresholding X_1 and X_2

1. Order the values for each variable in $\mathbf{X} = (X_1, X_2)$, to yield $oX = (oX_1, oX_2)$ which is the matrix $\mathbf{X}$ with values for X_1 and X_2 sorted in ascending order

2. Remove the lowest and highest 5% of values from the matrix $o\mathbf{X}$ from consideration

3. For each pair oX_{1_i}, oX_{2_j} where $i, j = 1, 2, \ldots, 0.9 * n$, calculate the cell counts for a 2x2 contingency table as follows:

$$\begin{aligned} a_{ij} &= \sum_{k=1}^{n} I(X_{1k} \geq oX_{1i}) \wedge I(X_{2k} \geq oX_{2j}) \wedge I(Y_k = 1) \\ b_{ij} &= \sum_{k=1}^{n} I(X_{1k} \geq oX_{1i}) \wedge I(X_{2k} \geq oX_{2j}) \wedge I(Y_k = 0) \\ c_{ij} &= \sum_{k=1}^{n} I(X_{1k} < oX_{1i}) \vee I(X_{2k} < oX_{2j}) \wedge I(Y_k = 1) \\ d_{ij} &= \sum_{k=1}^{n} I(X_{1k} < oX_{1i}) \vee I(X_{2k} < oX_{2j}) \wedge I(Y_k = 0) \end{aligned} \tag{3.6}$$

4. Select the pair (oX_{1_i}, oX_{2_j}) that maximizes the statistic $g(t|X_1, X_2)$, where $g(t)$ is one of the six statistics in Table 3.3. For example,

$$OR_{ij} = \frac{a_{ij} d_{ij}}{b_{ij} c_{ij}}$$

3.5 Simulation Study

In sections 3.2 and 3.3, we demonstrated that the six statistics defined in Table 3.3 are maximized at the true threshold **T** when response Y is associated with the continuous variables X_1 and X_2 through the relationship defined by Equation 3.1 whether X_1 and X_2 are dichotomized singly or jointly. Furthermore, we showed that joint dichotomization should converge to T_1, T_2 faster than single dichotomization for all six statistics if the relationship in Equation 3.1 is true. However, it is not generally known in advance whether or not Y is associated with two continuous variables independently or through their interaction. Therefore, we investigate the ability of joint and single thresholding to recover the true thresholds, T_1 and T_2, for two continuous variables, X_1 and X_2, to discriminate a binary outcome Y when X_1 and X_2 are associated with Y when sampling from a population. A simulation study was conducted to evaluate the ability of the six statistics to correctly find T_1 and T_2 under different scenarios arising from combinations of (1) the relationship between $X = (X_1, X_2)'$ and Y (independent or interaction), (2) strength of association between the predictors in X and response Y as defined by an odds ratio, and (3) value of the true thresholds T_1 and T_2.

Independent Case: We set $P_{Y=1}$, $P_{X_1} \geq T_1$, $P_{X_2} \geq T_2$, the odds ratio for $X_1 \geq T_1$, OR_1, and the odds ratio for $X_2 \geq T_2$, OR_2. In the case where the interaction, X_1X_2, is independently associated with Y, the OR is the product of OR_1 and OR_2. Continuous variables X_1 and X_2 are generated from $N_2 \sim (0, I_2)$ and T_1 and T_2 are defined based on $P_{X_1 \geq T_1}$ and $P_{X_2 \geq T_2}$. The four probabilities, $P_1 = P_{Y=1|X_1 \geq T_1, X_2 \geq T_2}, P_2 = P_{Y=1|X_1 \geq T_1, X_2 < T_2}, P_3 = P_{Y=1|X_1 < T_1, X_2 \geq T_2}, P_4 = P_{Y=1|X_1 < T_1, X_2 < T_2}$ can be calculated based on the set values of T_1, T_2, OR_1 and OR_2. Response Y is generated from $\text{Bin}(n, P_k), k = 1, \ldots, 4$ based on the observed values of X_1 and X_2. For the independent case, we consider the scenarios outlined in Table 3.4 where probabilities $P_{X_1 \geq T_1}$ and $P_{X_2 \geq T_2}$ of 0.05, 0.2, and 0.5 yield thresholds of 1.645, 0.84, and 0 respectively.

Joint case: We set $P_{X_1 \geq T_1}$, $P_{X_2 \geq T_2}$, $P_{Y=1}$, $P_{Y=1|X1 \geq T1, X2 \geq T2}$, and the OR for condition $X_1 \geq T_1$ and $X_2 \geq T_2$. Continuous variables X_1 and X_2 are generated from $N_2(0, I_2)$ and the true thresholds T_1 and T_2 are set as the inverse normal values of $P_{X_1 > T_1}$ and $P_{X_2 > T_2}$. Two probabilities $P_1 = P_{Y=1|X_1 \geq T_1, X_2 \geq T_2}$ and $P_2 = P_{Y=1|X_1 \geq T_1 \vee X_2 \geq T_2}$ are calculated from the set values of OR,

Table 3.4: Simulation Scenarios

OR	$P_{X_1 \geq T_1} = P_{X_2 \geq T_2}$	$P_{Y=1}$	Scenario
1.5	0.05	0.2	a
	0.2	0.2	b
	0.5	0.2	c
3	0.05	0.2	d
	0.2	0.2	e
	0.5	0.2	f
6	0.05	0.2	g
	0.2	0.2	h
	0.5	0.2	i

$P_{Y=1}$, $P_{X_1 \geq T_1}$ and $P_{X_2 \geq T_2}$, . Response Y is generated from Bin$(n, P_w), w = 1, \ldots, 2$ based on the observed values of X_1 and X_2. For the joint case, we consider the scenarios outlined in Table 3.4 where probabilities $P_{X_1 \geq T_1} = P_{X_2 \geq T_2}$ of 0.05, 0.2, and 0.5 yield thresholds of 1.645, 0.84, and 0 respectively.

For each simulation scenario outlined in Table 3.4, we generated 500 datasets of sample size 500. The threshold for each method was estimated using the single and joint thresholding algorithms described in Section 3. The ability of each method to recover the true thresholds, T_1 and T_2, was evaluated by examining the mean squared error and the bias squared for the estimated threshold across all simulated datasets for all scenarios. All simulations were conducted in R v. 3.2.1 [55].

3.5.1 Simulation Results

Figures 3.3 and 3.4 show the results for thresholding X_1 singly and jointly. The results for thresholding X_2 were similar.

3.5.2 Independent case

In the independent case, as the strength of association between X_1, X_2 and Y increases (OR=1.5 to OR=6), both joint and single thresholding exhibit smaller MSE for the estimated threshold for all methods and bias decreases slightly suggesting that the estimated threshold is less variable and biased as the strength of association between X_1, X_2 and Y increases. When the strength of association is ≥ 3, single thresholding is better for all methods except odds ratio and relative risk. Joint thresholding

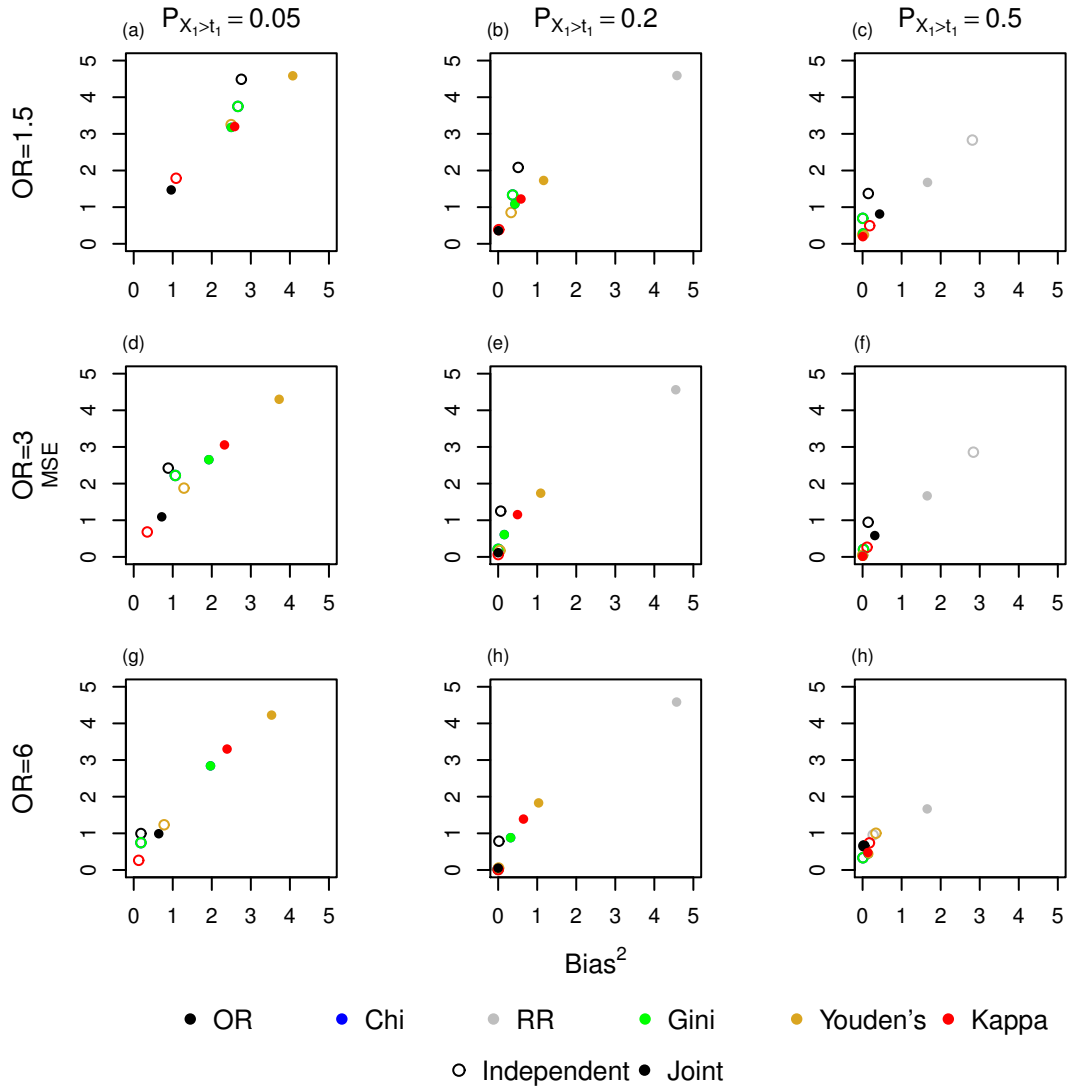

Figure 3.3: The results from the simulation study comparing joint and single dichotomization of independent continuous variables. Each graph shows the mean squared error (MSE) by bias squared for all statistics described in Table 3.3 for the different values of $P_{X_1 \geq T_1}$ and $P_{X_2 \geq T_2}$ and strength of association with Y. The columns show the impact of increasing values for $P_{X_1 \geq T_1}$ and $P_{X_2 \geq T_2}$ and the rows show the impact of increasing strength of association with Y. Filled circles represent joint thresholding while open circles represent single thresholding.

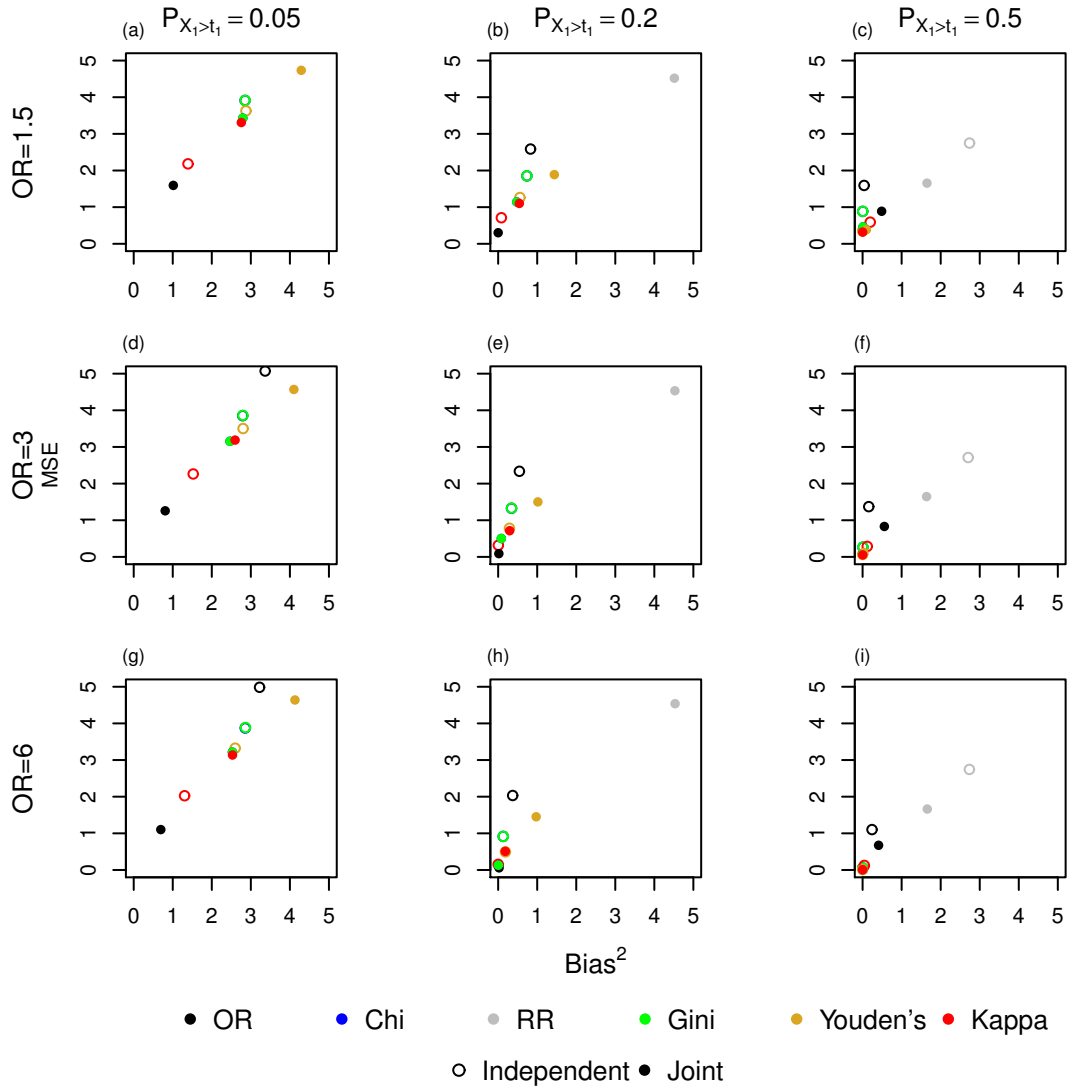

Figure 3.4: The results from the simulation study comparing joint and single dichotomization of interacting continuous variables. Each graph shows the mean squared error (MSE) by bias squared for all statistics described in Table 3.3 for the different values of $P_{X_1 \geq T_1}$ and $P_{X_2 \geq T_2}$ and strength of association with Y. The columns in Figure 3.4 show the impact of increasing values for $P_{X_1 \geq T_1}$ and $P_{X_2 \geq T_2}$ and the rows show the impact of increasing strength of association with Y. Filled circles represent joint thresholding while open circles represent single thresholding.

for kappa statistic has a lower MSE and bias than single thresholding for kappa when $P_{X_1 \geq T_1}$ and $P_{X_2 \geq T_2}$ increases to 0.5.

Holding odds ratio constant, as the probabilities $P_{X_1 \geq T_1}$ and $P_{X_2 \geq T_2}$ increase, both joint and single thresholding show a reduction in bias. At the lowest odds ratio (OR=1.5) as $P_{X_1 \geq T_1}$ and $P_{X_2 \geq T_2}$ increase, the threshold estimated jointly using odds ratio, Gini Index, or relative risk has lower MSE and bias relative to the threshold selected using single thresholding. However, the jointly estimated threshold using Youden's statistic or kappa statistic has higher bias and MSE than the threshold selected using single thresholding. When $P_{X_1 \geq T_1}$ and $P_{X_2 \geq T_2} = 0.5$ (Figure 3.4c), selecting a threshold jointly based on kappa improves relative to single thresholding. Relative risk has the highest MSE and bias for both joint and single thresholding. Relative risk is not shown for plots 3.4a,d, and g due to the magnitude of the MSE and bias.

3.5.3 Joint Case

In the joint case, as the probability of observing values of $P_{X_1 \geq T_1}$ and $P_{X_2 \geq T_2}$ increase, both joint and single thresholding show a reduction in MSE and bias. As was seen in the independent case, selecting a threshold jointly using odds ratio, Gini Index, or relative risk result in a lower MSE and bias than single thresholding. However, the jointly estimated threshold using Youden's statistic or kappa statistic has a higher MSE and bias than the threshold selected using single thresholding. When $P_{X_1 \geq T_1}$ and $P_{X_2 \geq T_2} = 0.05$ or 0.2, selecting a threshold singly using kappa has a lower MSE and bias than jointly. But as the probability increases to $P_{X_1 \geq T_1}$ and $P_{X_2 \geq T_2} = 0.5$, selecting a threshold jointly using kappa improves relative to single thresholding (Figures 3.4c,f, and i). When $P_{X_1 \geq T_1}$ and $P_{X_2 \geq T_2} = 0.5$, single and joint thresholding using Youden's statistic results in an MSE and bias approximately zero.

As the strength of association between X_1, X_2 and Y increases (OR=1.5 to OR=6), both joint and single thresholding exhibit a reduction in MSE and bias decreases slightly suggesting that the estimated threshold is less variable and biased as the strength of association increases. Selecting a threshold jointly using odds ratio results in the lowest MSE and bias of all the methods except when $P_{X_1 \geq T_1}$ and $P_{X_2 \geq T_2} = 0.5$. At this highest probability, selecting a threshold jointly and singly using

chi square, gini index, Youden's statistic and Kappa statistic result in a lower MSE and bias relative to joint thresholding using odds ratio. Single thresholding using relative risk results in the highest MSE and bias of all the methods.

3.5.4 Summary of Results

When X_1 and X_2 are independently associated with Y, single thresholding results in a lower MSE and bias when there is a weak association and small probability of observing values above a threshold. As that association and probability increase, joint thresholding performs similarly or better than single thresholding. When X_1 and X_2 are associated with Y described by an interaction as described by Equation 3.1, joint thresholding with the odds ratio method results in the lowest MSE and bias when there is a weak or modest association with response variable Y. When there is a strong association and a high probability of observing values above a certain threshold, all of the methods except relative risk yield a low MSE and bias for the estimated thresholds.

3.6 Conclusion

Dichotomizing variables is often necessary for many clinical and statistical purposes. Also, identifying interactions that lead to increased risk of disease is an important step in understanding disease etiology. If two or more variables are dichotomized independently, their association with the outcome may never be identified. Thus, if continuous variables must be dichotomized and there is a suspected interaction, joint dichotomization is ideal. Joint dichotomization could be thought of as a first step in possibly identifying these interactions in data.

This paper provided mathematical and numeric proof that if X_1 and X_2 are associated with outcome through an interaction, joint dichotomization (1) yields a larger statistic for odds ratio, relative risk, chi square, Youden's, Kappa and (2) converges more quickly to a true threshold T than single thresholding. Through a simulation study, we showed that when a binary outcome is associated with two continuous variables through an interaction, dichotomizing them jointly to discriminate Y recovers the true threshold with less variability than dichotomizing singly. Of the six statistics

investigated, simulations showed that maximizing the odds ratio provided the most improvement when dichotomizing jointly instead of singly.

One limitation of this paper is our choice of simulation data. We chose to simulate data that represents a step function relationship between X_1, X_2, and Y thus, there is a greater probability of observing $P_{Y=1}$ when both $P_{X_1 \geq T_1}$ and $P_{X_2 \geq T_2}$ than when they are not. This is possible when disease outcome can be described by a mixture of normal distributions meaning disease negative has one distribution and disease positive has another distribution. However, other relationships between interactions and disease may not be favorable for joint dichotomization.

In situations where interactions between variables are suspected and there is a need to dichotomize the continuous variables, these variables should be dichotomized jointly. However, our simulations showed that even in the independent case when X_1 and X_2 were associated with the outcome, joint thresholding was still shown to be effective in recovering a true threshold. In the case of the odds ratio statistic, joint thresholding performs better whether there is an interaction or not.

3.7 Chapter 3 Supplemental Material

The response surface for X_1, X_2, and each of the statistics based on joint dichotomization is rough and, therefore, we also considered applying the smoothing algorithm described in Box 3 to the matrix of statistics for each combination of thresholds for X_1, and X_2. We then choose the thresholds that yield the maximum statistic after smoothing.

Box 3: Smoothing Algorithm

1. On response surface S, define the *m* x *m* matrix, **M**, around each point $s = (oX_{1i}, oX_{2j}, OR_{ij})$.
2. Replace the value of OR_{ij} with the mean of **M**.
3. If a square of dimensions *m* cannot be formed around point s, reduce the size of the square to the largest possible rectangle, R, that does not exceed the area of **M**.

4. The X_1 and X_2 values that correspond with the absolute maximum of the surface are selected as the thresholds.

Figure 3.5 shows an example of the smoothing algorithm for 50 observations.

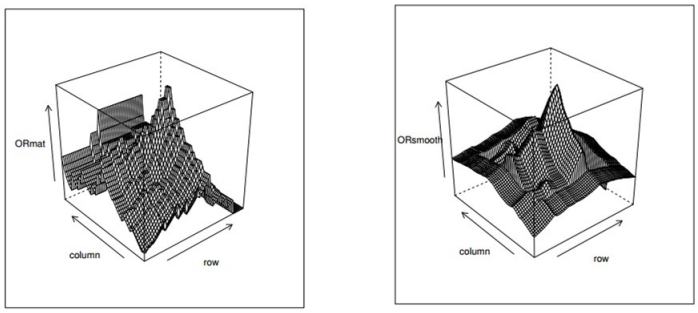

Figure 3.5: Odds Ratio surface before and after smoothing

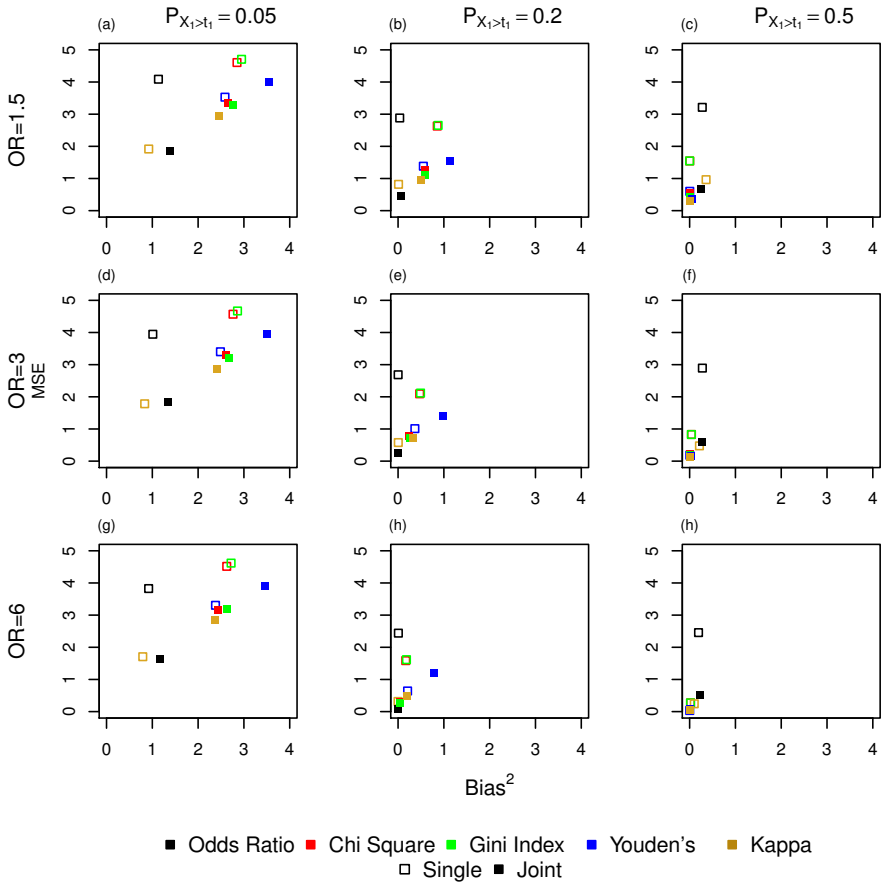

Figure 3.6: The results from the simulation study comparing joint and single dichotomization of interacting continuous variables *after smoothing*. Each graph shows the mean squared error (MSE) by bias squared for all statistics described in Table 3.3 for the different values of $P_{X_1 \geq T_1}$ and$P_{X_2 \geq T_2}$ and strength of association with Y. The columns in Figure 3.6 show the impact of increasing values for $P_{X_1 \geq T_1}$ and $P_{X_2 \geq T_2}$ and the rows show the impact of increasing strength of association with Y. Filled squares represent joint thresholding while open squares represent single thresholding.

3.7.1 Additional plots for Chapter 3

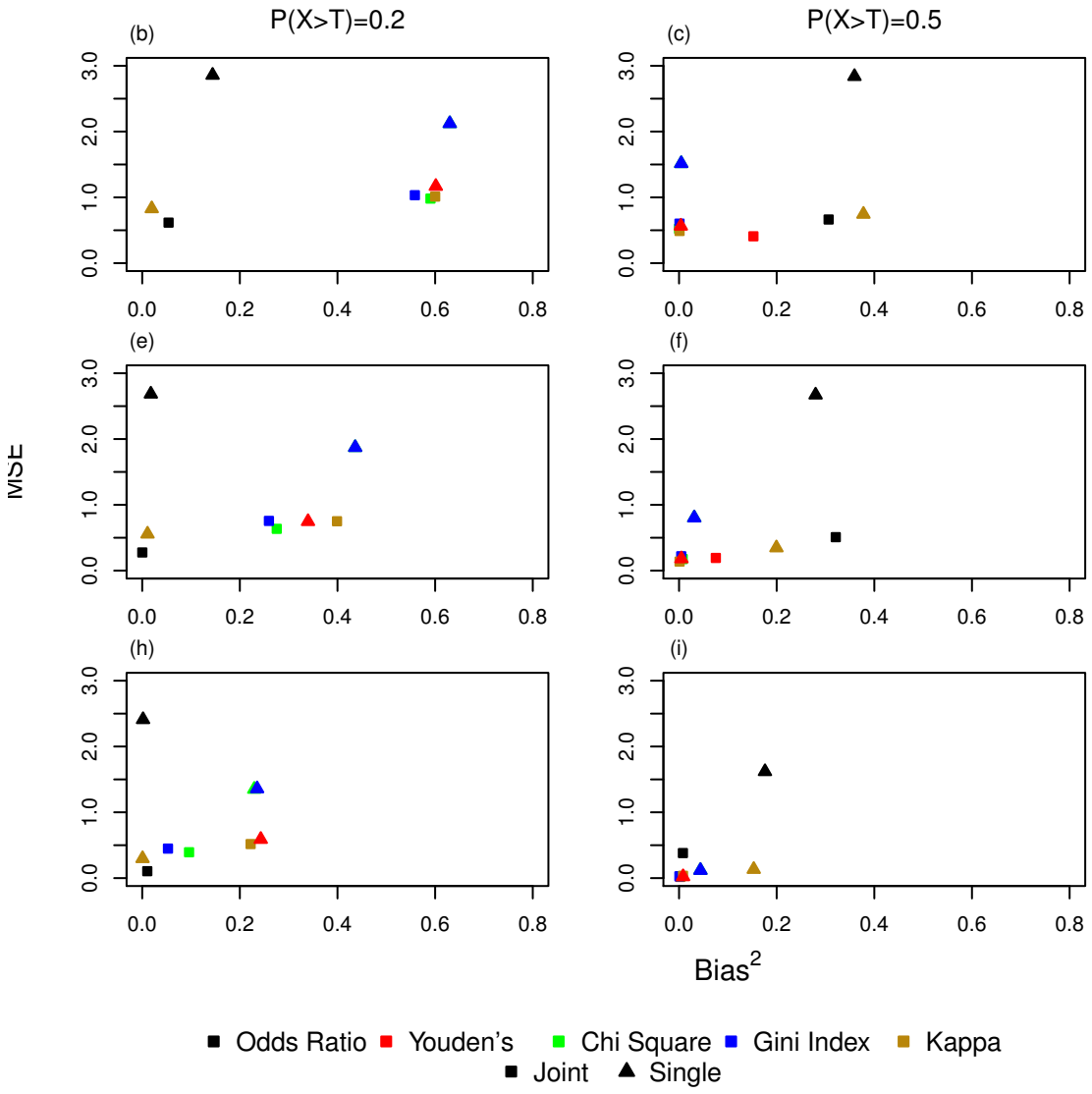

Figure 3.7: Joint versus single thresholding for $P(X_1 \geq T_1) = 0.2$ and $P(X_1 \geq T_1) = 0.5$

AN EXTENSION OF THE LOGIC REGRESSION FRAMEWORK FOR THE INCLUSION OF CONTINUOUS VARIABLES IN THE IDENTIFICATION OF INTERACTIONS THAT INCREASE RISK OF DISEASE

4.1 Introduction

An important goal in medicine is to understand the role of genetic variation and environmental factors in the context of disease risk in order to improve diagnosis, prevention and treatment [57]. Though much research has been done to identify genes and environmental factors that increase risk of disease, it has been suggested that only through the identification of statistical interactions between these genes and environment factors can additional progress be made to improve disease prediction [1]. A study by Aschardet et al found that even though many genetic and clinical risk factors likely to contribute to the etiology of complex diseases, they were unable to identify statistical interactions between these factors that lead to increased risk or disease prediction [58]. This could indicate that new statistical methods are needed to investigate these interactions. In a statistical interaction, the association of an effect measure (e.g. age) with outcome differs in the presence of a third variable (e.g. smoking). Thus, for example, the association between age and disease may not be detected unless age is paired with another factor such as smoking. This type of interaction may be difficult to detect with traditional statistical applications.

Logistic Regression is a common statistical approach often used to model dichotomous outcomes from continuous and binary data. When investigating interactions, however, these should be hypothesized *a priori*. When there are a small number of variables, all possible combinations of the variables can be included in the model without difficulty. For example, if there are 4 main effects, the number of total terms to include in the model would be $2^4 - 1 = 15$ (i.e. 4 main effects, 6 two-way interactions, 4 three-way interactions, and 1 four-way interaction). If the number of variables increases to 20, however, the number of model terms becomes $2^{20} - 1 = 1,048,575$ which is not feasible. Also, when the number of parameters exceeds the number of observations, which

is common in genetic data, then any logistic regression will be overspecified resulting in too many variables in the model and increasing the risk of collinearity.

Machine Learning techniques proved an alternative to logistic regression and there is no need to identify interactions *a priori*. Artificial neural networks (ANN) can model complex data, but the models are a "Black Box" which is difficult to interpret. Support Vector Machines (SVMs), another machine learning technique can classify disease outcome, but like ANN the results often lack interpretability. These methods focus on prediction and not the identification of interactions. Once again, for SVM, if the number of variables exceeds the number of observations, the method can also result in overspecification.

Tree-based methods are easily interpreted statistical models based on how factors are associated with the outcome. They can model a dichotomous outcome like traditional statistical models, but interactions do not have to be determined *a priori*. Unlike SVM and ANN, decision trees focus on classification in addition to prediction and they can still perform if the number of parameters exceed the number of observations. Thus, they will be the focus of this chapter.

This chapter develops an alternative decision tree algorithm for identifying continuous and binary interactions associated with disease. Section 4.2 of this paper will explore current tree-based methods for interaction identification as well as propose an alternative method. Section 4.3 will discuss a simulation designed to explore how this new methodology compares to current practices while Section 4.4 will present the results. Section 4.5 will provide a discussion and review of the topics present as well as make recommendations for use of new methodology.

4.2 Current Methods

As described in Chapter 1, CART is a common decision tree method that splits data into increasingly homogeneous groups until a stopping rule is applied. Figure 4.1 gives an example of a simple CART tree built from three binary predictors. This tree predicts an individual to be in category 0 if $X_1 = 0$ or $X_1 = 1, X_2 = 0$ and $X_3 = 0$. It predicts category 1 if $X_1 = 1$ and $X_2 = 1$, or if $X_1 = 1, X_2 = 0$ and $X_3 = 1$.

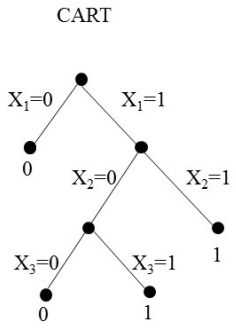

Figure 4.1: A CART tree with three binary predictors (X_1, X_2, and X_3) and two classes (0 and 1).

CART can be used for binary and continuous data types, but as stated in Chapter 1, CART is biased toward the inclusion of continuous variables meaning that if it has a choice of including a continuous variable or a binary variable into the model, it will choose the continuous [2]. Also, by design, once CART makes a split on a variable, a branch is formed and certain combinations of variables are no longer viable.

Logic regression is an alternative tree based method that uses Boolean logic to model a binary outcome. The use of Boolean logic allows it to form specific combinations of variables that can not be formed with CART. All CART tree models can be written with a Boolean logic statement but not all Boolean logic statements can be represented in a CART tree. For example, consider the logic statement $D = 1 : (X_1 \wedge X_2) \vee (X_1 \wedge X_3)$ which means an interaction between variables X_1 and X_2 or an interaction between X_1 and X_3 leads to disease. As shown in figure 4.2, this can be modeled exactly with logic regression but not with CART.

Logic regression, however, is currently not designed for the inclusion of continuous variables. Thus, we endeavor to find a way to split continuous variables within logic regression. To do this, we

explore the dichotomization methods described in Chapter 2 of this dissertation and shown in Table 4.1.

Odds Ratio	Youden's Statistic	Chi-Square
$\frac{ad}{bc}$	$\frac{a}{a+c} + \frac{d}{b+d} - 1$	$\frac{(ad-bc)^2}{(a+b)(c+d)(b+d)(a+c)}$
Kappa Statistic	Relative Risk*	Gini Index
$\frac{(a+d)-((a+b)(a+c)+(c+d)(b+d))}{1-((a+b)(a+c)+(c+d)(b+d))}$	$\frac{a/(a+b)}{c/(c+d)}$	$(P_y(1-P_y)) - (\frac{ab}{a+b} + \frac{cd}{c+d})$

Table 4.1: Formulas for statistics for selecting a threshold for a continuous variable X to discriminate a binary outcome Y based on the probabilities in a 2X2 contingency table *For cohort study

4.2.1 Subset Matching

Let Q be a set of a interaction terms (e.g. $X_1 \wedge X_2 \wedge X_3$) identified by a decision tree method consisting of T trees where $D = X_1 \wedge X_3$ is the exact interaction of interest. The set of terms Q is called a subset match because $D \in Q$. D is also a subset match because $D \in D$. Figure 4.2 shows how both CART and logic regression model this interaction. Logic regression can model interaction D exactly while CART can only achieve a subset match Q.

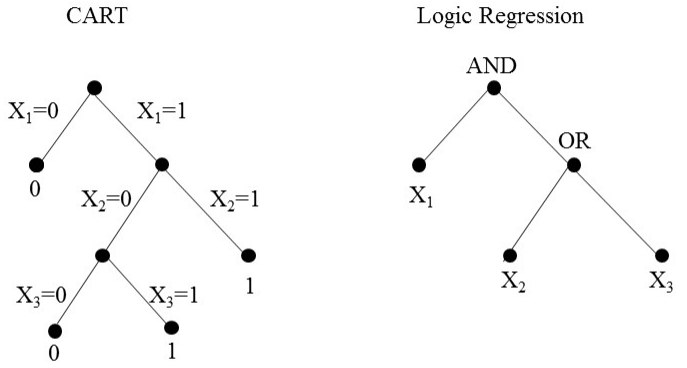

Figure 4.2: Comparison of a CART tree and a logic regression tree. Logic regression models $(X_1 \wedge X_2) \vee (X_1 \wedge X_3)$ which includes the exact match to $D = X_1 \wedge X_3$ while CART must model $(X_1 \wedge X_2) \vee (X_1 \wedge X_2 \wedge X_3)$ which includes the subset match to D, $Q = X_1 \wedge X_2 \wedge X_3$

If a method repeatedly detects an interaction of interest with added extraneous variables (i.e. subset matching), it may lead to erroneous conclusions on the data. Thus, understanding the benefit of exact matching over subset matching will aid in the analysis of the results.

4.2.2 C.Logic Algorithm

Algorithm for C.Logic

1. For each continuous variable, X, in each sample order the values to create a variable oX
2. For each continuous variable in each sample calculate the cell counts for a $2x2$ contingency

table as follows:

$$
\begin{aligned}
a_k &= \sum_{k=1}^{t} I(X_k \geq oX_k) \wedge I(Y_k = 1) \\
b_k &= \sum_{k=1}^{t} I(X_k \geq oX_k) \wedge I(Y_k = 0) \\
c_k &= \sum_{k=1}^{t} I(X_k < oX_k) \wedge I(Y_k = 1) \\
d_k &= \sum_{k=1}^{t} I(X_k < oX_k) \wedge I(Y_k = 0)
\end{aligned}
\tag{4.1}
$$

3. Separate the candidate predictor variables into binary and continuous

4. For each continuous variable, at each value of oX_k calculate the a new threshold by using on of the statistics in Table 4.1

5. Find the maximum value of each statistic from Table 4.1. The oX value that corresponds to the maximum value is the new threshold.

6. Dichotomize the continuous variables of the original data set with the new thresholds.

7. Combine the original binary values that were separated in step 3 with the newly formed binary values

8. This new data set is used in logic regression.

4.3 Simulation Study

For the simulation, we generate three binary variables, X_1, X_2 and X_3, under a Bernoulli distribution with probability of 0.3. Next we generate X_4, X_5 and X_6 under a multivariate distribution such that $\mathbf{X} \sim N_3(\mathbf{0}, I_3)$. We set true thresholds for continuous variables T_4, T_5, and T_6, as the inverse normal value of $P_{X_i} \geq 0.3$ for $i = 4, \ldots 6$ which is $T_i = 0.524$ for $i = 4, \ldots 6$. Let $L = (X_1 \wedge X_2) \vee (X_3 \wedge X_4) \vee (X_5 \wedge X_6)$ and $P(Y = 1) = P(L = 1)P(Y = 1|L = 1) + P(L = 0)P(Y = 1|L = 0)$ with $P(Y = 1|L = 1) > P(Y = 1|L = 0)$ Finally, we generate 11 binary noise variables, $\mathbf{X} \sim B(0, 0.5)$ and 3 continuous noise variables,$\mathbf{X} \sim N_3(\mathbf{0}, I_3)$. For our simulation, we use sample sizes of 200,

300, 500, 1000, 1500, and 2000. We also consider odds ratios between our binary outcome Y and logic statement L of 2,4, and 8. In logic regression, we use starting and ending temperatures of 3 and -1 respectively with 100000 iterations as the parameters for the simulated annealing algorithm. THe starting and ending temperatures were selected such that >90% of worse models would be selected at the initiation of the annealing chain and <1% of worse models were accepted at the finial iteration of the annealing chain. As discussed in Chapter 1, the starting and ending temperatures indicate where in the annealing chain logic regression is and thus adjusts the probability of accepting new moves. For additional details about simulated annealing see Ruczinski et al [6]. Each simulation is run with 500 repetitions.

In section 4.2, we presented an algorithm, C.Logic, for extending logic regression to include continuous variables. In this section, we compare C.Logic to CART in order to investigate which method performs better in identifying interactions that are associated with the outcome. First, we investigate which of the statistics from Table 4.1 perform best in the C.Logic algorithm. Figure 4.3 shows a plot of the number of times the interaction of interest was correctly identified at sample sizes 200, 300, 500, 1000, 1500, and 2000 and for an odds ratio of 4.

4.4 Simulation Results

First, we investigate which of the five methods - odds ratio, Youden's statistic, chi square statistic, gini index, kappa statistic - perform best in the C.Logic algorithm. We eliminated relative risk from consideration since it performed so poorly in the previous chapters. Figure 4.3 shows a comparison of each of the five methods along with CART. For this simulation, we looked at an OR=4, $P_{Y=1} = 0.2$, and $P(Y = 1|L = 1) = 0.304$. We also looked at subset matching and exact matching.

4.4.1 Evaluation of choice of statistic

Subset Matching Since X_1X_2 is already binary, we were not concerned with how well the statistics performed in identifying that interaction. Instead we focused on X_3X_4 and X_5X_6. For the binary/continuous interaction X_3X_4, CART recovers the interaction more often than all the methods

until sample size 1500 at which point Kappa and Youden's find the interaction at a higher rate than CART. Odds ratio performs poorly and only identifies this interaction about 20% of the time. For the continuous/continuous interaction, X_5X_6, CART finds the interaction more often than all the methods. Youden's and Kappa improve as sample size increases and performs comparably to CART at sample size 2000.

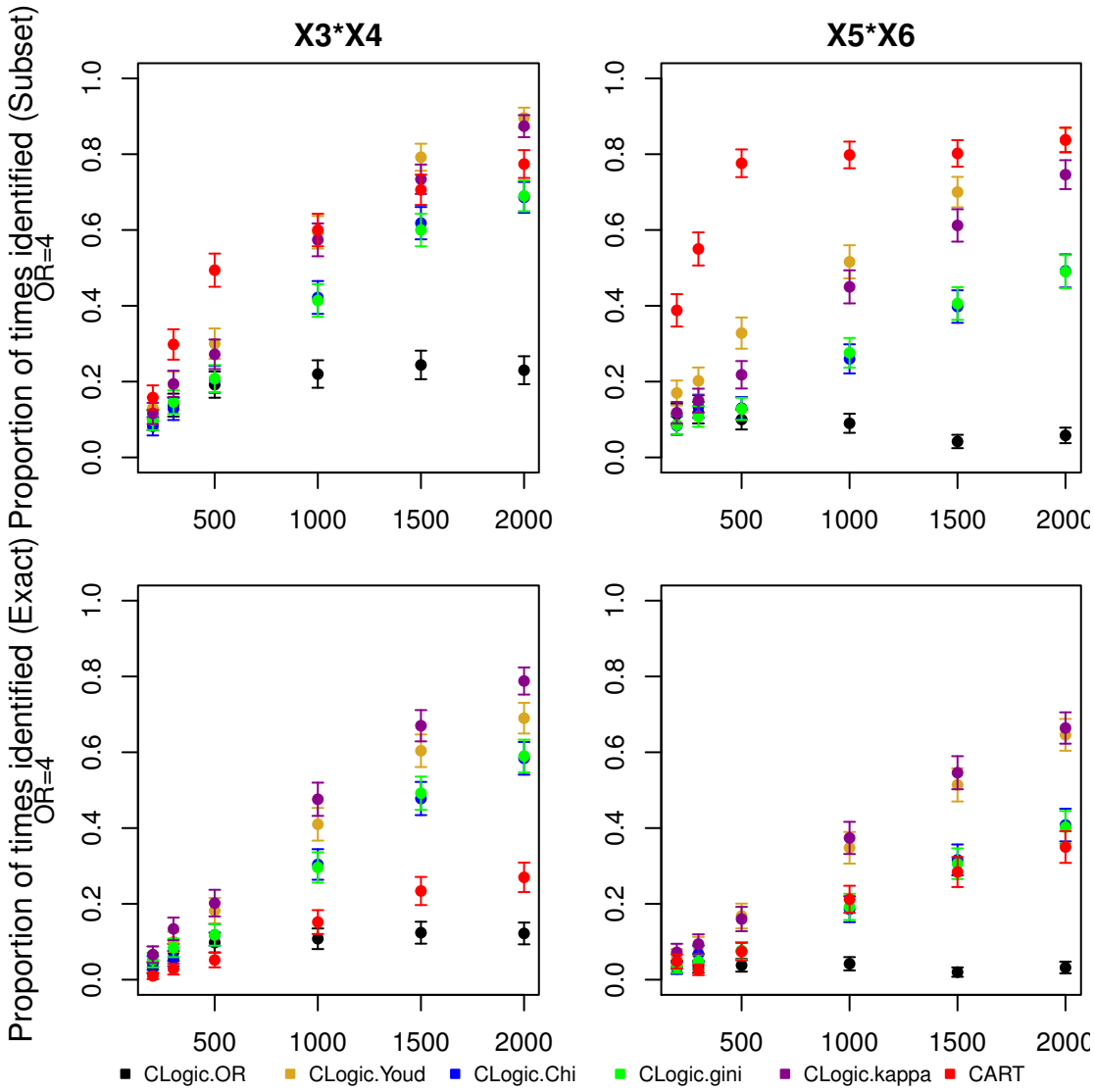

Figure 4.3: Simulation results for C.Logic using all methods except relative risk compared to CART

Exact Matching For the binary/continuous interaction X_3X_4 and the continuous/continuous interac-

tion X_5X_6, all five methods except odds ratio recover the threshold at a higher rate than CART. Of the five methods, Kappa performs the best in all the scenarios followed closely by Youden's. Next, we use kappa in the C.Logic algorithm as the default statistic.

4.4.2 Comparison of CART and C.Logic

Subset Matching Figure 4.4 shows that for all odds ratios, C.Logic with the kappa statistic identifies the X_1X_2 interaction more often than CART. For the binary/continuous interaction X_3X_4, CART recovers the interaction more often than C.Logic until sample size 1500 and odds ratio 4 at which point C.Logic identifies the interaction at a higher rate than CART. CART recovers the X_5X_6 interaction more often than C.Logic at all odds ratio. The tendency for CART to do better with continuous variables than binary variables is to be expected since, as stated in Chapter 1, CART is biased toward the inclusion of continuous variables [2].

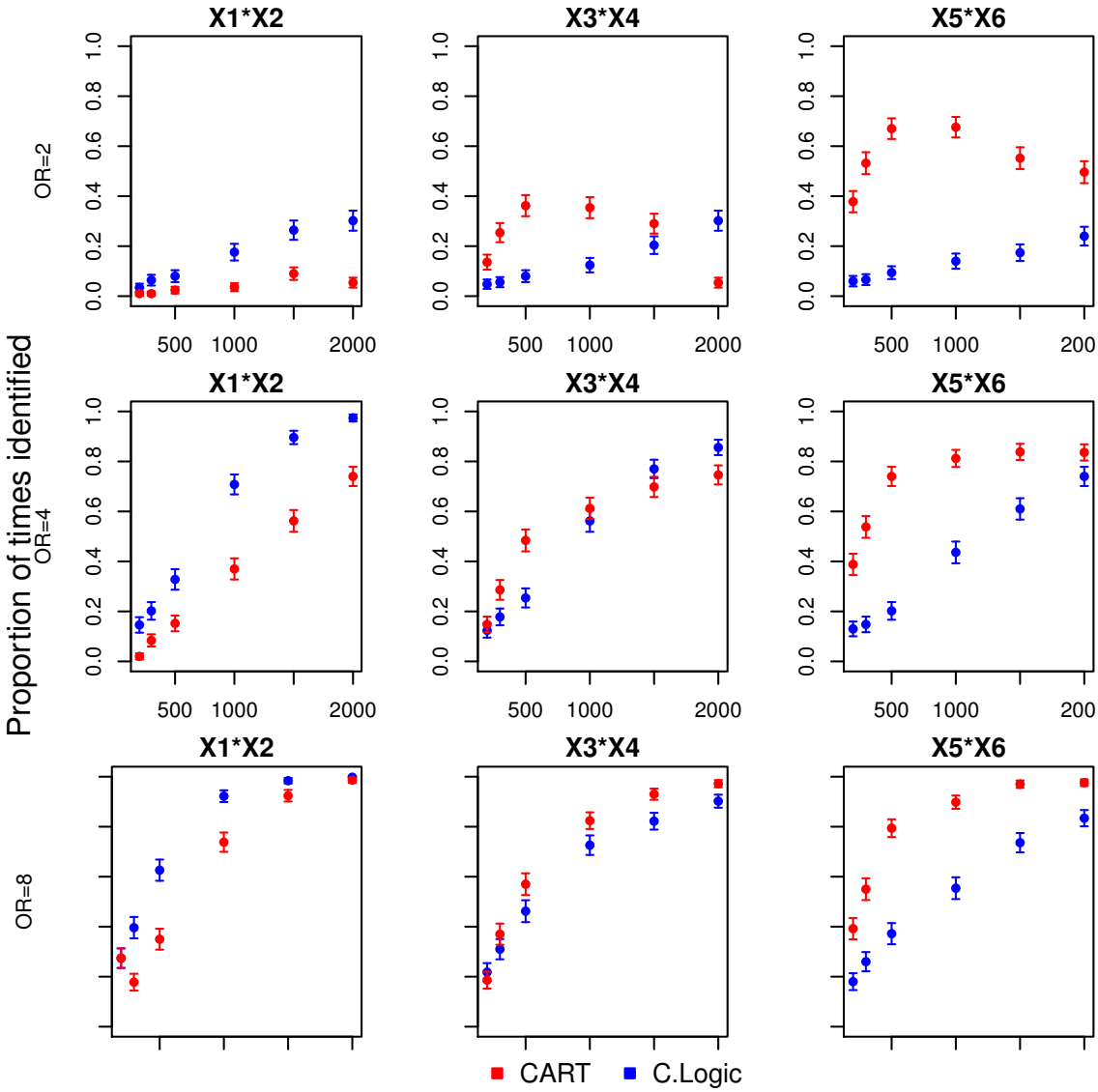

Figure 4.4: Simulation results for C.Logic using Kappa compared to CART

Exact Matching When looking at exact matching in Figure 4.5, the benefit of C.Logic is clear. C.Logic exactly recovers the interactions of interest more often than CART at all odds ratios. The greatest improvement can be found in the continuous/continuous interaction. In subset matching, CART identified this interaction more often than C.Logic at every odds ratio and sample size even reaching 100% at odds ratio 8. In exact matching, however, it never identifies the interaction more than 30% of the time.

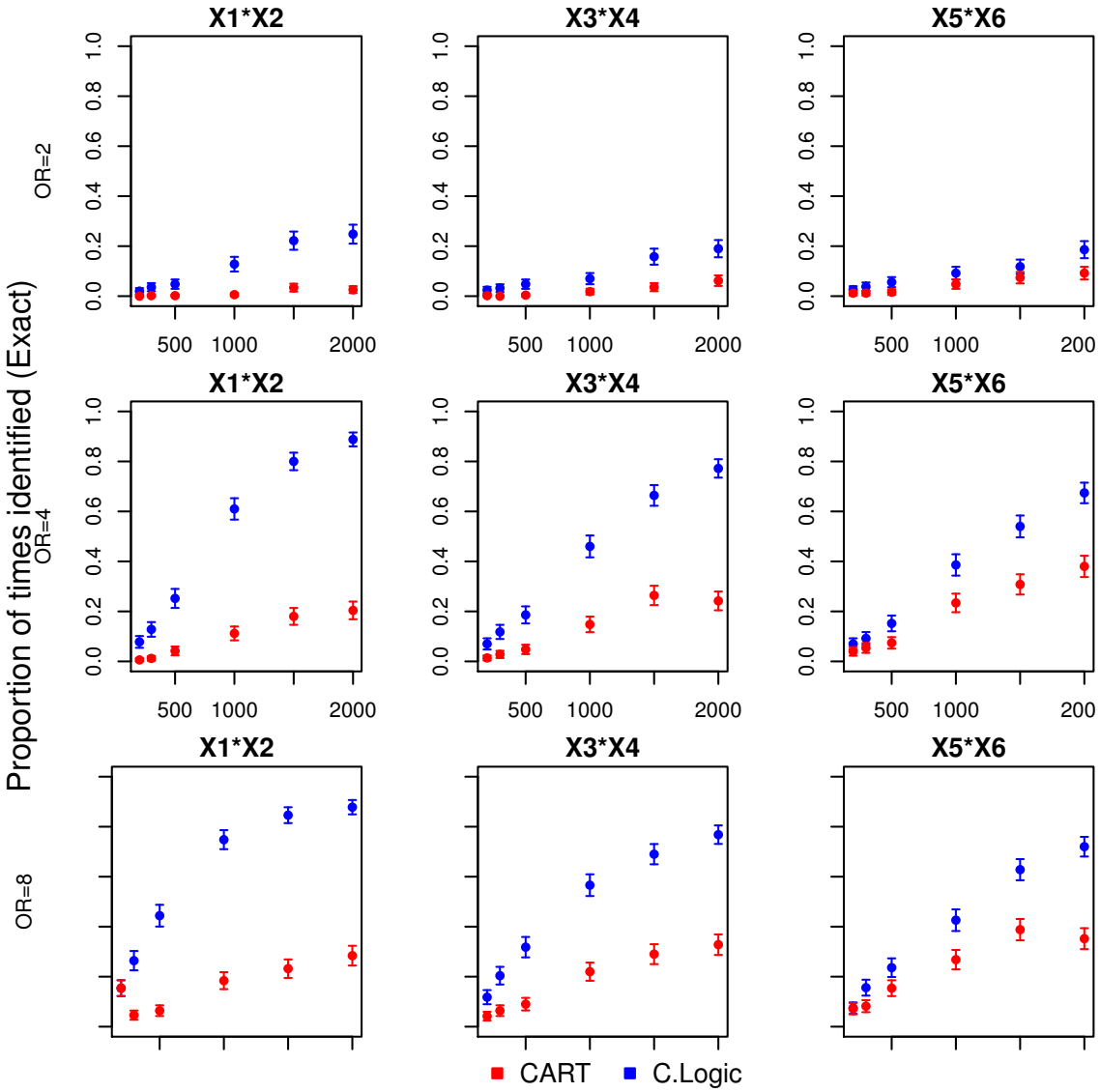

Figure 4.5: Simulation results for C.Logic using Kappa compared to CART

4.5 Application

4.5.1 Periodontis in African Americans with Diabetes

In this section, we use C.Logic and CART to examine the association between adult periodontal disease and certain genetic and environmental factors. The data comes from a study conducted at the Center for Oral Health Research at the Medical University of South Carolina and defines periodontal

disease as $\geq$ 3mm clinical attachment loss at 30% of affected sites. The periodontal disease data set consists of 244 African Americans (AA) with type 2 diabetes mellitus. The factors studied include 14 continuous or categorical health indicators for total cholesterol, HDL, triglycerides, C-reactive protein levels, HbA1c levels, BMI, flossing, dental visits, insurance, age, income, brushing, sex and smoking status as well as 19 SNPs believed to play a role in inflammation and/or bone resorption. Of the 244 participants 89 had periodontal disease and 155 did not.

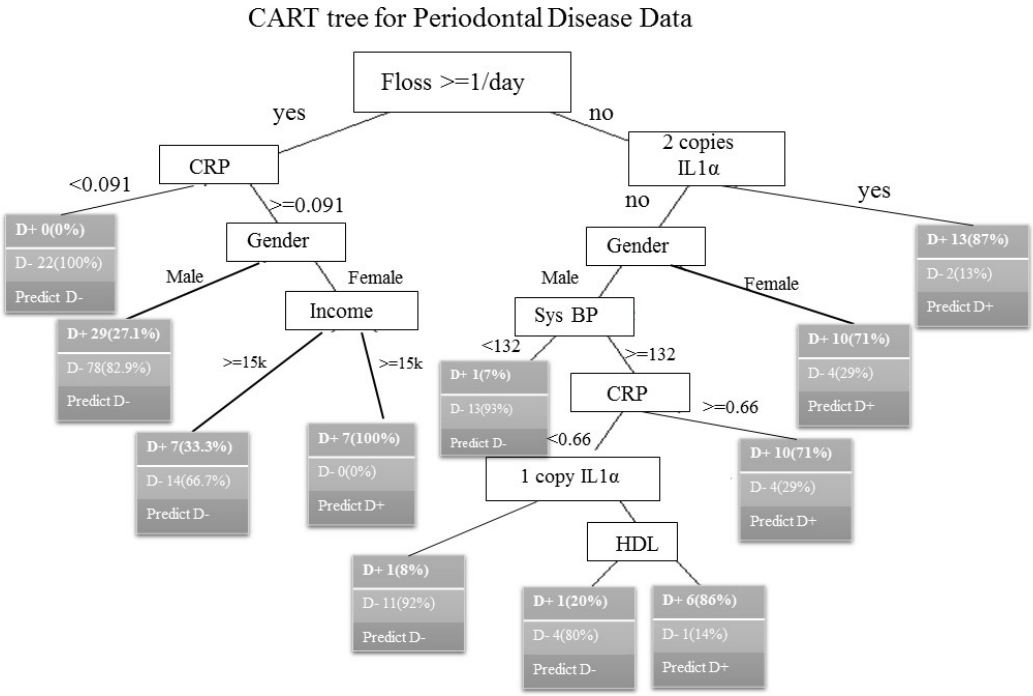

Figure 4.6: CART model results for Periodontal Disease data

Both CART and C.Logic found an association between being female and increased risk of disease. For CART, being female along with income $\geq$ 15k, c-reactive protein $\geq$ 0.091 and flossing each day lead to increased risk of periodontal disease, as well as being female with two copies of the risk allele in *IL1*α gene and not flossing each day. In the C.Logic model, being female interacted with SBP $\geq$ 146, DBP $\geq$ 81 and HBa1c > 7%. Systolic blood pressure and *IL1*α were also found in both models. More investigation would be needed to evaluate the importance of these variables in increased risk of

periodontal diseases in AAs with diabetes.

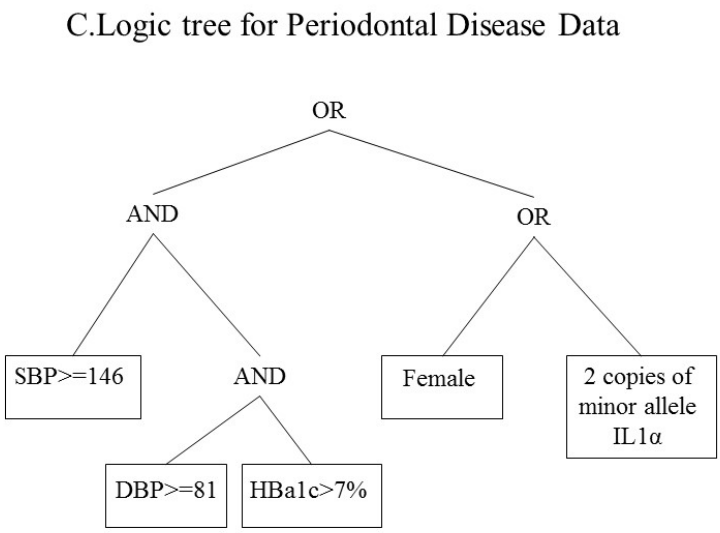

Figure 4.7: C.Logic model results for Periodontal Disease data. L=L=((SBP $\wedge$ DBP $\wedge$ HBa1c) $\vee$ ((Female $\vee$ *IL1*α)))

4.5.2 Biomarkers in Lupus Nephritis (LN)

As discussed in Chapter 1, LN is one of the most common complications associated with SLE affecting approximately 50% of SLE patients [15]. This data set includes urine biomarkers from 140 patients with biopsy proven LN recruited by the MUSC Division of Rheumatology and Immunology Lupus Erythematosus clinical research group (MUSCLE) to a study whose goal was to develop models of LN outcome from a novel collection of urine biomarkers. Biomarkers used for analysis include interleukin 6(*IL6*), interferon inducible protein 10(IP10), eotaxin-1, granulocyte macrophage stimulating factor (GMCSF), interferon $\alpha 2$ (*IFN*$\alpha 2$), interferon γ (IFNγ), interleukin -1α (*IL1*α), IL1β, monocyte chemo attractant protein-1 (MCP1), MIP-1, platelete-derived growth factor BB

chain (PDGF-BB), Lipocalin 2 (NGAL), osteoprotegerin (OPG), cystatin c (CysC), and N-acetyl-beta-D-glucosaminidase (NAG).

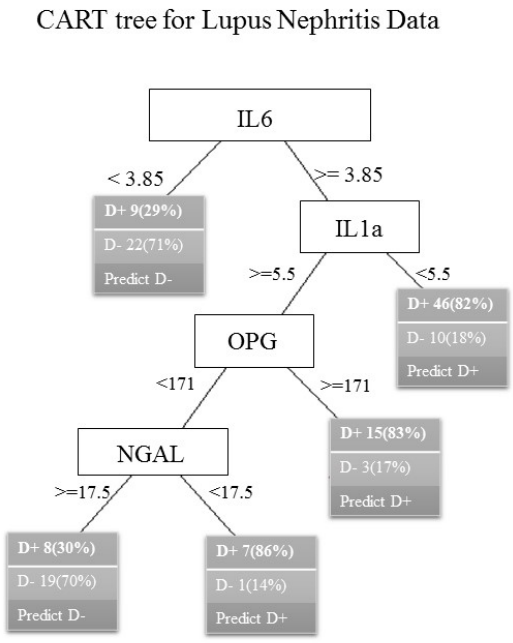

Figure 4.8: CART model results for Lupus Nephritis data

The CART model created its first split at 3.85 for *IL6*. Figure 4.8 shows that CART found interactions that lead to increased risk of disease between (*IL6* ≥ 3.85 and *IL1*$\alpha < 5.5$), ($ILg \geq 3.85, IL1a < 5.5$ and $OPG \geq 171$), and (*IL6* ≥ 3.85, *IL1*$\alpha < 5.5, OPG \geq 171$ and $NGAL < 17.5$).

C.Logic found interactions including $L = MIP1B \geq 9.2 \wedge EGFR \geq 90.7 \vee IL6 \geq 42.4 \wedge MCP1 \geq 183.4 \vee IL1a \geq 22.6 \wedge MCP1 \geq 183.4$. The thresholds chosen by CART and C.Logic for the IL6 variable (3.6 and 42.6, respectively) are rather different. There are no other variables common between the methods. Further investigation would have to be done to evaluate the importance of these variables.

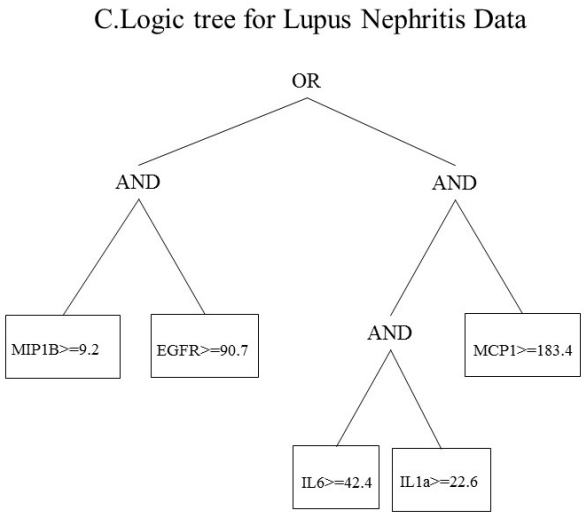

Figure 4.9: C.Logic results for Lupus Nephritis data. L=((MIP-1α $\wedge$ EGFR) $\vee$ ((IL6 $\vee$ Il1α) $\wedge$ MCP1))

4.6 Conclusion

Identifying interactions that lead to increased risk of disease is an important problem in both medicine and statistics. If a factor increases risk of disease only in the presence of another factor, it may go undiscovered with traditional statistical methods. CART is a tree-based method that can identify interactions to model a binary outcome. Studies have shown, however, that CART can be biased toward the inclusion of continuous variables [2]. Logic regression is specifically designed to find interactions that are associated with outcome but it is not designed for continuous variables. In this paper, we created a new algorithm, C.Logic, that allows for the inclusion of continuous variables in a logic regression framework. C.Logic uses five methods of dichotomization previously shown to successfully recover a true threshold. With the exception of odds ratio, any of the methods perform better than CART at exactly identifying interactions associated with outcome. The default method, kappa statistic, exactly identifies the interaction of interest more often than CART for every sample size and strength of association.

CART is superior to our method C.Logic in subset matching of the interactions of interest. That is to say that if X_3X_4 is the true interaction that causes increased risk of disease, CART will identify this interaction but along with other variables that are not associated with outcome. Thus, CART is more sensitive while C.Logic is more specific. If recovering true relationships between interactions and outcomes is the goal, C.Logic is the preferred method.

The application of the CART and C.Logic algorithms to the periodontal disease data set found a few variables in common (i.e female, IL1α, and systolic blood pressure). The application to the Lupus Nephritis data set found the *IL6* variable in common. In general, the C.Logic trees are simpler to interpret. Both models provide a basis for further investigation of possible variables of interest. Though CART results are not to be discounted, based on the results of the simulation studies, the C.Logic results are more likely to reflect true interactions found in the data.

4.7 Chapter 4 Supplemental Material

Algorithm for MedC.Logic

1. Take t bootstrap samples of the data
2. For each continuous variables, X, in each sample order the values to create a variable oX
3. For each continuous variable in each sample calculate the cell counts for a $2x2$ contingency table as follows:

$$\begin{aligned} a_k &= \sum_{k=1}^{t} I(X_k \geq oX_k) \wedge I(Y_k = 1) \\ b_k &= \sum_{k=1}^{t} I(X_k \geq oX_k) \wedge I(Y_k = 0) \\ c_k &= \sum_{k=1}^{t} I(X_k < oX_k) \wedge I(Y_k = 1) \\ d_k &= \sum_{k=1}^{t} I(X_k < oX_k) \wedge I(Y_k = 0) \end{aligned} \tag{4.2}$$

4. At each value of X_t, use the corresponding contingency table to calculate the six statistics

found in Table 4.1

5. Find the maximum value for each statistic for each variable in each bootstrap sample.

6. For each continuous variable, find the median of the bootstrap samples. These medians are the new thresholds.

7. Dichotomize the continuous variables of the original data set with the new thresholds.

8. Apply logic regression

Figure 4.10 shows the results from a cohort study design scenario. Each plot shows the sample size by the proportion of times the interaction is correctly identified exactly. The columns in the figure show the interactions: X_1X_2, X_3X_4, and X_5X_6, as described in section 3. The rows show the impact of increasing the strength of association with Y, represented by the odds ratio in the data. For each interaction, as the sample size increases both C.Logic and CART improve in identifying the interactions. As the odds ratio increases so does the proportion of times the interactions are exactly identified in a C.Logic or CART tree. Both methods perform poorly when identifying the X_5X_6 interaction when the odds ratio is 2. As the odds ratio increases to 8, however, C.Logic improves and performs better than CART with C.Logic identifying the correct interactions 75% of the time compared to 37% with CART. C.Logic outperforms CART by the greatest margin for the X_1X_2 interaction when odds ratio is 8.

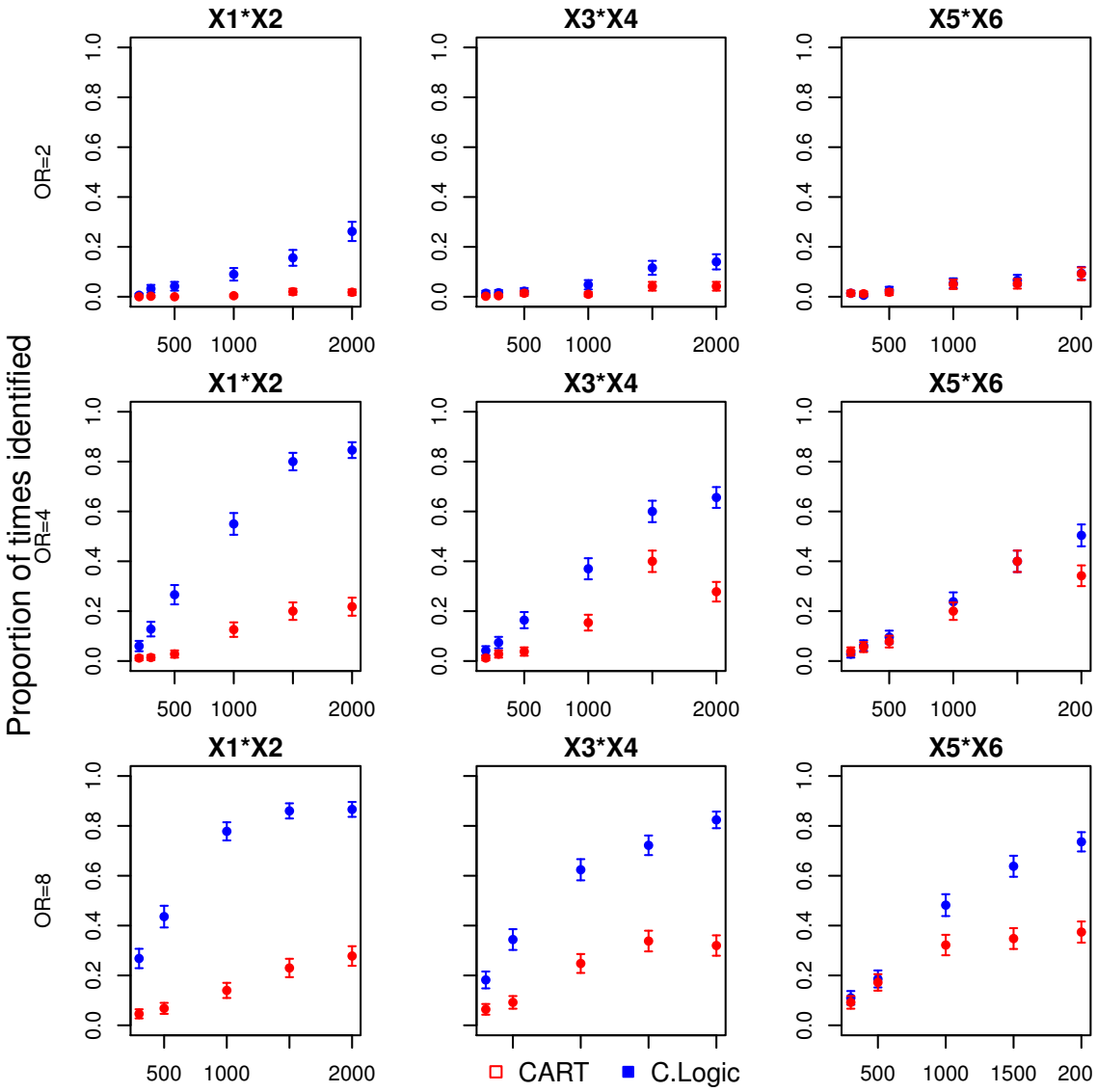

Figure 4.10: Simulation results showing sample size by the proportion of times the interactions are exactly identified under the case-control study design comparing MedC.Logic to CART with bootstrapping. MedC.Logic takes the median threshold value of the five statistics.

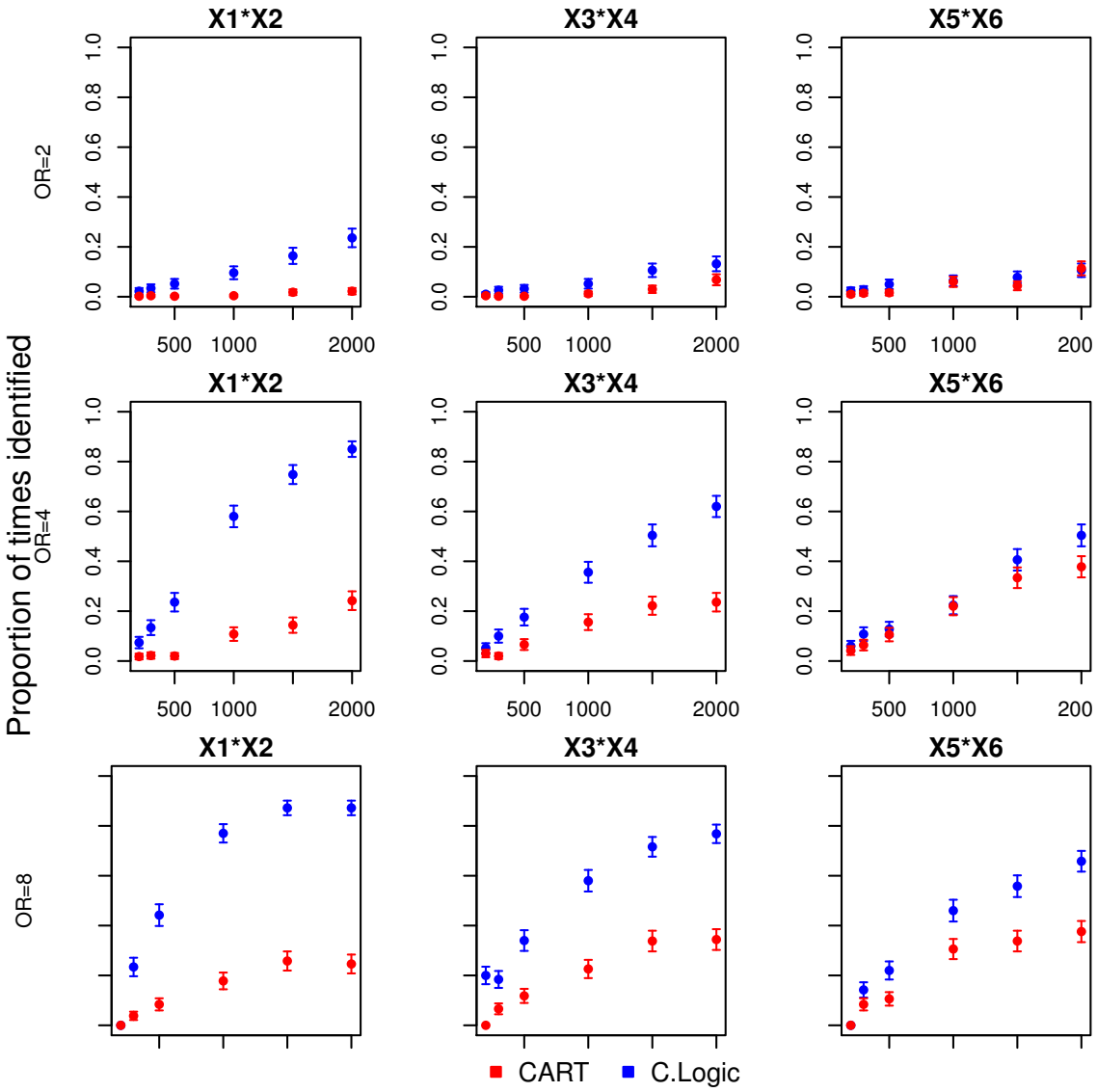

Figure 4.11: Simulation results showing sample size by the proportion of times the interactions are identified exactly or in a subset with other factors under the case-control study design comparing MedC.Logic to CART without bootstrapping.

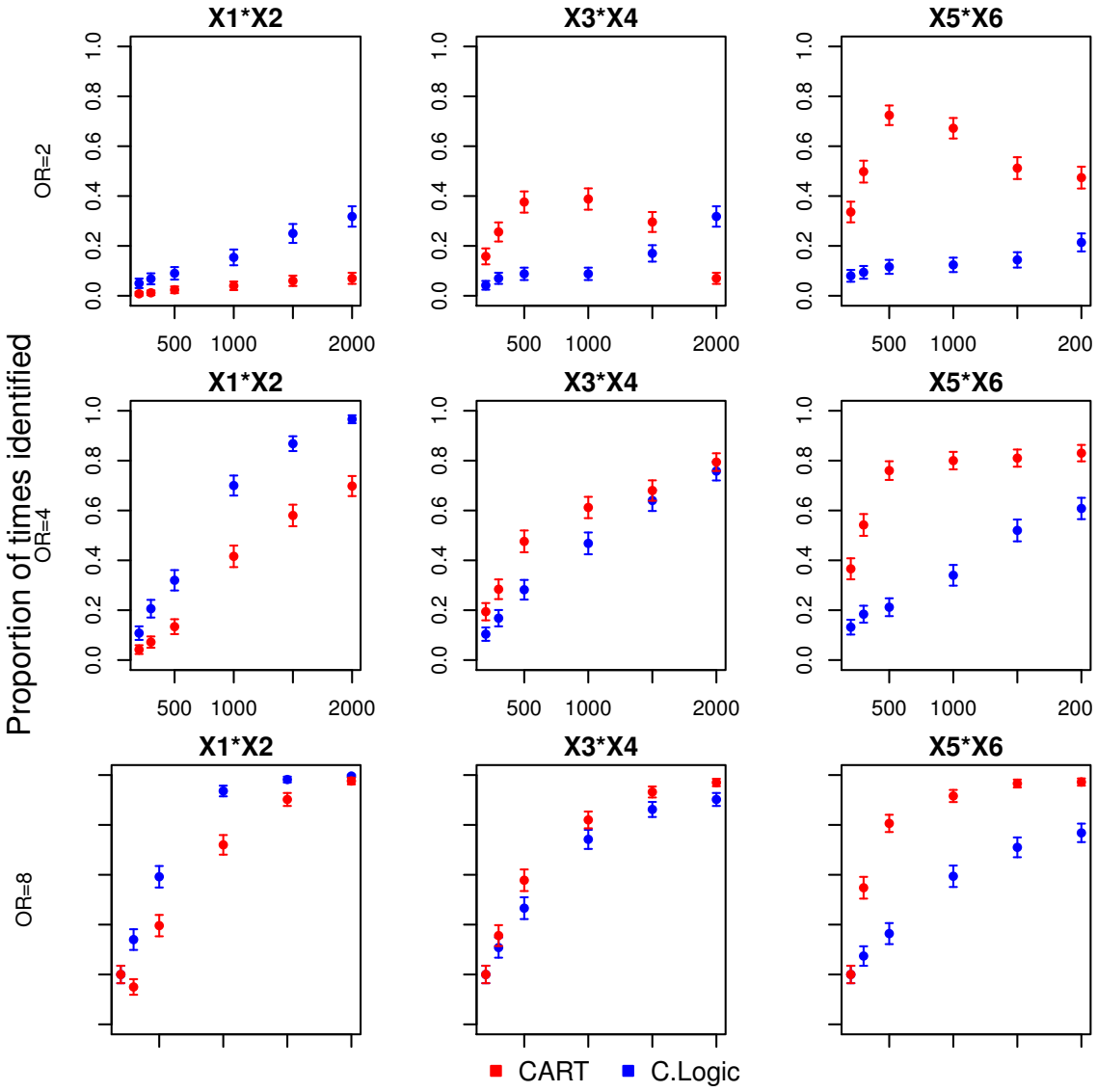

Figure 4.12: Simulation results showing sample size by the proportion of times the interactions are identified in a subset with other factors under the case-control study design comparing C.Logic to CART

CONCLUSION

5.1 Summary

The motivation for this project was the identification of gene-gene and gene-environment interactions associated with increased risk of disease. Many diseases have a complex etiology, characterized by a combination of genetic and environmental factors, yet, common practice is to disregard main effects that are not associated with outcome before considering whether the factor is associated in the presence of another factor. In other words, it is possible that the interaction of two factors leads to increased risk of disease without the main effect of the individual factors.

Because many clinical measurements such as cholesterol level or blood pressure are continuous, many clinicians dichotomize before analysis or treatment recommendations. Thus, proper dichotomization of factors is a first step to understanding the relationship between variables.

Clinicians dichotomize for a variety of reasons such as risk stratification, drug dosage or treatment recommendations. Statisticians often dichotomize for ease of interpretation or to implement certain statistical methods. With many methods of dichotomization available, it is important to know which ones are effective. This project showed that for our specific set of parameters, maximizing certain statistics is the most appropriate for dichotomization. We showed theoretically and numerically that maximizing odds ratio, relative risk, Youden's, Chi Square, Gini Index, and Kappa are the best ways to recover a true threshold given one exists. A simulation study confirmed that these six statistics recover the threshold and showed that increasing the odds ratio and choice of threshold improves these statistics ability to recover the true threshold.

Furthering the discussion of dichotomization, we also considered the dichotomization of interactions. If two factors are interacting to increase risk of disease, yet they are dichotomized independently, their relationship to each other and to the outcome could become obscured. In this project, we created an algorithm to dichotomize one variable while considering its relationship with another variable. To do this, we plotted the six statistics mentioned in the previous section for every

possible combination of two variables and identified the absolute maximum. The values of the factors at this maximum were chosen as the thresholds. In the supplemental materials, we also explored adding a smoothing algorithm and bootstrapping to this process in order to improve the selection of a threshold.

After simulation studies, we found that for strong odds ratios, joint dichotomization recovered the true threshold better than dichotomizing variables individually. It was interesting to note that in some cases, even where there was no interaction between the variables, dichotomizing jointly was still better than dichotomizing singly.

Next we focused on identifying interactions in a data set. There are many methods for identifying interactions of variables that are associated with outcome. Many of these methods require that the interactions be identified *a priori* (e.g logistic regression) or they determine thresholds sequentially as opposed to simultaneously (e.g. CART). Logic regression is a tree based method specifically designed for the identification of interactions but it does not allow for continuous variables. The C.Logic algorithm extends the logic regression framework for the conclusion of continuous variables. Based on the study of dichotomization methods above, it uses five of the dichotomizations, along with Boolean logic in order to correctly identify interactions that lead to increased risk of disease. One can choose which of the methods to use in the algorithm or use kappa which has been set as the default since it performed the better than the other methods. In the supplement material, we also explored the performance of C.Logic when using the mean of the five methods of dichotomization in order to select the thresholds.

Simulation studies show that C.Logic performs better than CART in exactly identifying interactions that lead to increased risk of disease. The simulation study in Chapter 4 shows that CART is superior to C.Logic in identifying subset matches of the interactions under investigation. That is to say, if X_1X_2 is the interaction of interest, CART will include that interaction in a branch but along with other variables that are actually not associated with the outcome. Thus, if recovering true relationships between factors and a binary outcome is the goal, C.Logic would be recommended over CART.

5.2 Limitations

This project only considered a specific structure of data. We designed the data in the simulations so that it was a mixture of binomials which followed a steep sigmoidal relationship between Y and the independent variable or interaction. Thus, the probability of $Y = 1$ is very small for small values of X or the interaction term. Then, at a certain point (i.e. the true threshold), the probability of observing $Y = 1$ increases rapidly. A data population similar to this construction would perform well with the methods described in this project. Other types of data construction would require further investigation.

5.3 Future Directions

One possible future direction is to rewrite the logic regression package so that it includes another step for the dichotomization process. This additional step could be the one responsible for selected the best threshold value before moving on to the Boolean logic phase of the method. This additional step could also allow for multiple choice of threshold on a single variable thereby letting the method adjust until the optimal threshold is found.

Within the logic regression framework, we could also allow for the selection of multiple cutpoints on a single variable. This would essentially allow logic regression to convert a continuous variable into a categorical variable instead of simply a dichotomous variable.

Another future step for this line of research includes perhaps using a forest method approach to increase predictive ability. The C.Logic algorithm creates one tree to classify disease. With a forest approach we can bootstrap the data and apply C.Logic to each set thereby creating several trees or a "forest." We then gather the trees and use a voting process to determine which variables belong in a final model. The use of multiple trees as opposed to one should theoretically improve the efficiency of the model.

A possible future application of this methodology could be found in personalized medicine. This algorithm allows one to separate a heterogeneous group of patients into more and more homogeneous groups based on genetic, clinical and environmental factors thus streamlining and

improving diagnoses and treatment options.

APPENDICES

6.1 A: Proofs for Chapter 2

A.1*Important terms*

For the theoretical investigation of dichotomization methods, we considered a true threshold of X called T such that $P_{Y=1|X\geq T} > P_{Y=1|X<T}$. For each possible threshold chosen, t_x, there are three possibilities: $t_x < T$, $t_x = T$, and $t_x > T$. Each possible threshold, t_x creates new cell values of a $2X2$ contingency table.

1. $t_x < T$

$$\begin{aligned} a &= P_{X\geq T}P_{Y=1|X\geq T} + (P_{X<T} - P_{X<t_x})P_{Y=1|X<T} \\ b &= P_{X\geq t_x} - (P_{X\geq T}P_{Y=1|X\geq T} - (P_{X<T} - P_{X<t_x})P_{Y=1|X<T}) \\ c &= (P_{X<t_x})P_{Y=1|X<T} \\ d &= (P_{X<t_x}) - (P_{X<t_x})P_{Y=1|X<T} \end{aligned} \tag{6.1}$$

2. $t_x = T$,

$$\begin{aligned} a &= P_{X\geq T}P_{Y=1|X\geq T} \\ b &= P_{X\geq T} - P_{X\geq T}P_{Y=1|X\geq T} \\ c &= P_{X\leq T}P_{Y=1|X<T} \\ d &= P_{X\leq T} - P_{X<T}P_{Y=1|X<T} \end{aligned} \tag{6.2}$$

3. $t_x > T$

$$\begin{aligned} a &= P_{X\geq t_x}P_{Y=1|X\geq T} \\ b &= P_{X\geq t_x} - P_{X\geq t_x}P_{Y=1|X\geq T} \\ c &= P_{X<T}P_{Y=1|X<T} + (P_{X<t_x} - P_{X<T})P_{Y=1|X\geq T} \\ d &= (P_{X<t_x} - (P_{X<T}P_{Y=1|X<T} - (P_{X<t_x} - P_{X<T})P_{Y=1|X\geq T}) \end{aligned} \tag{6.3}$$

A.2 *Proof of Theorem 1 for Youden's Statistic*

Let X be a random variable and Y a dichotomous variable. Also, let T be a threshold such that, $P_{Y=1|X\geq T} > P_{Y=1|X<T}$. There are three possibe cases that can occur when selecting a threshold for X, t_x: (1) $t_x < T$,(2) $t_x = T$, and (3) $t_x > T$. The expression for the Youden's Statistic, $\frac{a}{a+c} + \frac{d}{b+d} - 1$, for each case can be found using the expressions for a,b,c, and defined in equations 6.1, 6.2, and 6.3. We can then show that the Youden's Statistic is maximized when $t_x = T$.

A.2.a Consider the case where $P_{X>t_x} > P_{X>T}$. Start with what is given

$$P_{Y=1|X\geq T} > P_{Y=1|X<T}$$

Multiply both sides by $P_{X\geq T}$

$$P_{X\geq T}P_{Y=1|X\geq T} > P_{X\geq T}P_{Y=1|X<T}$$

On the right hand side, let $P_{X\geq T}$ =1 $- P_{X<T}$

$$P_{X\geq T}P_{Y=1|X\geq T} > (1 - P_{X<T})P_{Y=1|X<T}$$

Add $P_{X<T}P_{Y=1|X<T}$ to both sides

$$P_{X\geq T}P_{Y=1|X\geq T} + P_{X<T}P_{Y=1|X<T} > P_{Y=1|X<T}$$

On the left hand side let $P_{X\geq T}P_{Y=1|X\geq T} + P_{X<T}P_{Y=1|X<T} = P_{Y=1}$

$$P_{Y=1} > P_{Y=1|X<T}$$

Multiply by $P_{t_x<X<T}$

$$P_{t_x<X<T}P_{Y=1} > P_{t_x<X<T}P_{Y=1|X<T}$$

Note $P_{t_x<X<T}$=$P_{X<T} - P_{x<t_x}$

$$P_{X<T}P_{Y=1} - P_{x<t_x}P_{Y=1} > P_{t_x<X<T}P_{Y=1|X<T}$$

Add $P_{x<t_x}P_{Y=1}$ to both sides

$$P_{X<T}P_{Y=1} > P_{t_x<X<T}P_{Y=1|X<T} + P_{x<t_x}P_{Y=1}$$

Subtract $P_{X<T}P_{Y=1}P_{Y=1|X<T}$ from both sides

$$\begin{aligned}
&P_{X<T}P_{Y=1} - P_{X<T}P_{Y=1}P_{Y=1|X<T} \\
&> P_{t_x<X<T}P_{Y=1|X<T} + P_{x<t_x}P_{Y=1} - P_{X<T}P_{Y=1}P_{Y=1|X<T}
\end{aligned}$$

On the right hand side, split $P_{X<T}P_{Y=1}P_{Y=1|X<T}$ into $P_{x<t_x}P_{Y=1}P_{Y=1|X<T}+P_{t_x<X<T}P_{Y=1}P_{Y=1|X<T}$

$$\begin{aligned}
&P_{X<T}P_{Y=1} - P_{X<T}P_{Y=1}P_{Y=1|X<T} > P_{t_x<X<T}P_{Y=1|X<T} \\
&+ P_{x<t_x}P_{Y=1} - P_{x<t_x}P_{Y=1}P_{Y=1|X<T} - P_{t_x<X<T}P_{Y=1}P_{Y=1|X<T}
\end{aligned}$$

Add $(1 - P_{Y=1})P_{X\geq T}P_{Y=1|x\geq T}$ to both sides

$$\begin{aligned}
&(1 - P_{Y=1})P_{X\geq T}P_{Y=1|x\geq T}P_{X<T}P_{Y=1} - P_{X<T}P_{Y=1}P_{Y=1|X<T} \\
&> (1 - P_{Y=1})P_{X\geq T}P_{Y=1|x\geq T}P_{t_x<X<T}P_{Y=1|X<T} \\
&+ P_{x<t_x}P_{Y=1} - P_{x<t_x}P_{Y=1}P_{Y=1|X<T} - P_{t_x<X<T}P_{Y=1}P_{Y=1|X<T}
\end{aligned}$$

Factor both sides

$$\begin{aligned}
&(1 - P_{Y=1})P_{X\geq T}P_{Y=1|x\geq T}P_{Y=1}(P_{X<T} - P_{X<T}P_{Y=1|X<T}) \\
&> (1 - P_{Y=1})(P_{X\geq T}P_{Y=1|x\geq T} + P_{t_x<X<T}P_{Y=1|X<T}) \\
&+ P_{Y=1}(P_{x<t_x}(1 - P_{Y=1|X<T}))
\end{aligned}$$

Divide both sides by $(1 - P_{Y+1})$ and $P_{Y=1}$

$$\frac{(1-P_{Y=1})P_{X\geq T}P_{Y=1|x\geq T}P_{Y=1}(P_{X<T}-P_{X<T}P_{Y=1|X<T})}{(1-P_{Y+1})P_{Y=1}}$$
$$> \frac{(1-P_{Y=1})(P_{X\geq T}P_{Y=1|x\geq T}+P_{t_x<X<T}P_{Y=1|X<T})+P_{Y=1}(P_{x<t_x}(1-P_{Y=1|X<T}))}{(1-P_{Y+1})P_{Y=1}}$$

Simplify

$$\frac{P_{X\geq T}P_{Y=1|x\geq T}}{P_{Y=1}}+\frac{(P_{X<T}-P_{X<T}P_{Y=1|X<T})}{(1-P_{Y=1})}$$
$$> \frac{(P_{X\geq T}P_{Y=1|x\geq T}+P_{t_x<X<T}P_{Y=1|X<T})}{P_{Y=1}}$$
$$+\frac{P_{x<t_x}(1-P_{Y=1|X<T})}{(1-P_{Y=1})}$$

Subtract 1 from both sides

$$\frac{P_{X\geq T}P_{Y=1|x\geq T}}{P_{Y=1}}+\frac{(P_{X<T}-P_{X<T}P_{Y=1|X<T})}{(1-P_{Y=1})}-1$$
$$> \frac{(P_{X\geq T}P_{Y=1|x\geq T}+P_{t_x<X<T}P_{Y=1|X<T})}{P_{Y=1}}$$
$$+\frac{P_{x<t_x}(1-P_{Y=1|X<T})}{(1-P_{Y=1})}-1$$

We have

$$\frac{a_{t_x=T}}{(a+c)_{t_x=T}}+\frac{d_{t_x=T}}{(b+d)_{t_x=T}}-1>\frac{a_{t_x<T}}{(a+c)_{t_x<T}}+\frac{d_{t_x<T}}{(b+d)_{t_x<T}}-1$$

Thus, $Youd_{t_x=T} > Youd_{t_x<T}$.

A.2.b Now consider the case where $P_{X>t_x} < P_{X>T}$. Start with what is given

$$P_{Y=1|X\geq T} > P_{Y=1|X<T}$$

Multiply both sides by $P_{X<T}$

$$P_{X<T}P_{Y=1|X\geq T} > P_{X<T}P_{Y=1|X<T}$$

On the left, let $P_{X<T} = (1 - P_{Y=1|X\geq T})$

$$P_{Y=1|X\geq T} - P_{X>T}P_{Y=1|X\geq T} > P_{X<T}P_{Y=1|X<T}$$

Add $P_{X>T}P_{Y=1|X\geq T}$ to both sides

$$P_{Y=1|X\geq T} > P_{X<T}P_{Y=1|X<T} + P_{X>T}P_{Y=1|X\geq T}$$

Note $P_{Y=1} = P_{X<T}P_{Y=1|X<T} + P_{X>T}P_{Y=1|X\geq T}$

$$P_{Y=1|X\geq T} > P_{Y=1}$$

Multiply both sides by $P_{T<X<t_x}$

$$P_{T<X<t_x}P_{Y=1|X\geq T} > P_{T<X<t_x}P_{Y=1}$$

Subtract $P_{T<X<t_x}P_{Y=1}P_{Y=1|X\geq T}$ from both sides

$$P_{T<X<t_x}P_{Y=1|X\geq T} - P_{T<X<t_x}P_{Y=1}P_{Y=1|X\geq T} > P_{T<X<t_x}P_{Y=1} - P_{T<X<t_x}P_{Y=1}P_{Y=1|X\geq T}$$

Let $P_{T<X<t_x} = P_{X\geq T} - P_{X\geq t_x}$

$$(P_{X\geq T}-P_{X\geq t_x})P_{Y=1|X\geq T}-(P_{X\geq T}-P_{X\geq t_x})P_{Y=1}P_{Y=1|X\geq T} > P_{T<X<t_x}P_{Y=1}-P_{T<X<t_x}P_{Y=1}P_{Y=1|X\geq T}$$

Distribute

$$\begin{aligned}
&P_{X\geq T}P_{Y=1|X\geq T} - P_{X\geq t_x}P_{Y=1|X\geq T} \\
&- P_{X\geq T}P_{Y=1}P_{Y=1|X\geq T} + P_{X\geq t_x}P_{Y=1}P_{Y=1|X\geq T} \\
&> P_{T<X<t_x}P_{Y=1} - P_{T<X<t_x}P_{Y=1}P_{Y=1|X\geq T}
\end{aligned}$$

Add $P_{X\geq t_x}P_{Y=1|X\geq T}$ to both sides and subtract $P_{X\geq t_x}P_{Y=1}P_{Y=1|X\geq T}$ from both sides

$$P_{X\geq T}P_{Y=1|X\geq T} - P_{X\geq T}P_{Y=1}P_{Y=1|X\geq T}$$
$$> P_{X\geq t_x}P_{Y=1|X\geq T} - P_{X\geq t_x}P_{Y=1}P_{Y=1|X\geq T} + P_{T<X<t_x}P_{Y=1} - P_{T<X<t_x}P_{Y=1}P_{Y=1|X\geq T}$$

Note $P_{T<X<t_x}P_{Y=1} = (P_{X<t_x} - P_{X<T})P_{Y=1}$ and add $P_{X<T}P_{Y=1}$ to both sides

$$P_{X\geq T}P_{Y=1|X\geq T} - P_{X\geq T}P_{Y=1}P_{Y=1|X\geq T} + P_{X<T}P_{Y=1}$$
$$> P_{X\geq t_x}P_{Y=1|X\geq T} - P_{X\geq t_x}P_{Y=1}P_{Y=1|X\geq T} + P_{X<t_x}P_{Y=1} - P_{T<X<t_x}P_{Y=1}P_{Y=1|X\geq T}$$

Subtract $P_{X<T}P_{Y=1}P_{Y=1|X<T}$ from both sides

$$P_{X\geq T}P_{Y=1|X\geq T} - P_{X\geq T}P_{Y=1}P_{Y=1|X\geq T} + P_{X<T}P_{Y=1} - P_{X<T}P_{Y=1}P_{Y=1|X<T}$$
$$> P_{X\geq t_x}P_{Y=1|X\geq T} - P_{X\geq t_x}P_{Y=1}P_{Y=1|X\geq T}$$
$$+ P_{X<t_x}P_{Y=1} - P_{X<T}P_{Y=1}P_{Y=1|X<T} - P_{T<X<t_x}P_{Y=1}P_{Y=1|X\geq T}$$

Divide both sides by $P_{Y=1}(1 - P_{Y=1})$ and factor

$$\frac{P_{X\geq T}P_{Y=1|X\geq T}}{P_{Y=1}} + \frac{P_{X<T}(1 - P_{Y=1|X<T})}{(1 - P_{Y=1})} > \frac{P_{X\geq t_x}P_{Y=1|X\geq T}}{P_{Y=1}}$$
$$+ \frac{P_{X<t_x} - P_{X<T}P_{Y=1|X<T} - P_{T<X<t_x}P_{Y=1|X\geq T}}{(1 - P_{Y=1})}$$

We have

$$\frac{a_{t_x=T}}{(a+c)_{t_x=T}} + \frac{d_{t_x=T}}{(b+d)_{t_x=T}} - 1 > \frac{a_{t_x>T}}{(a+c)_{t_x>T}} + \frac{d_{t_x>T}}{(b+d)_{t_x>T}} - 1$$

Thus, $Youd_{t_x=t} > Youd_{t_x>t}$. If the expression for $Youd_{t_x=T}$ is greater than the expression for $Youd_{t_x<T}$ and the expression for $Youd_{t_x=T}$ is greater than the expression for $Youd_{t_x>T}$ then it shows that the Youden's Statistic is the highest when $t_x = T$.

A.3 *Proof of Theorem 1 for Gini Index*

Let X be a random variable and Y a dichotomous variable. Also, let T be a threshold such that, $P_{Y=1|X\geq T} > P_{Y=1|X<T}$. There are three possibe cases that can occur when selecting a threshold for X, t_x: (1) $t_x < T$,(2) $t_x = T$, and (3) $t_x > T$. The expression for the Gini Index, $(P_y(1-P_y)) - (\frac{ab}{a+b} + \frac{cd}{c+d})$, for each case can be found using the expressions for a,b,c, and defined in equations 6.1, 6.2, and 6.3. We can then show that the Gini Index is maximized when $t_x = T$.

A.3.a Consider the case where $P_{X>t_x} > P_{X>T}$. Start with what is given

$$P_{Y=1|X\geq T} > P_{Y=1|X<T}$$

Subtract $P_{Y=1|X<T}$ from both sides

$$0 < P_{Y=1|X\geq T} - P_{Y=1|X<T}$$

Square both sides

$$0 < (P_{Y=1|X\geq T} - P_{Y=1|X<T})^2$$

Multiply

$$0 < P^2_{Y=1|X\geq T} - 2P_{Y=1|X\geq T}P_{Y=1|X<T} + P^2_{Y=1|X<T}$$

Subtract $P^2_{Y=1|X\geq T}$ and $P^2_{Y=1|X<T}$ from both sides

$$-P^2_{Y=1|X\geq T} - P^2_{Y=1|X<T} < -2P_{Y=1|X\geq T}P_{Y=1|X<T}$$

Multiply by $P_{X\geq T}P_{t_x<X<T}$

$$(-P_{t_x<X<T})P_{X\geq T}P^2_{Y=1|X\geq T} + (-P_{X\geq T})P_{t_x<X<T}P^2_{Y=1|X<T} < -2P_{X\geq T}P_{Y=1|X\geq T}P_{t_x<X<T}P_{Y=1|X<T}$$

Note $-P_{t_x<X<T} = P_{X\geq T} - P_{X>t_x}$ and $-P_{X\geq T} = P_{t_x<X<T} - P_{X>t_x}$

$$\begin{aligned}&(P_{X\geq T} - P_{X>t_x})P_{X\geq T}P_{Y=1|X\geq T}^2 + (P_{t_x<X<T} - P_{X>t_x})P_{t_x<X<T}P_{Y=1|X<T}^2\\ <& -2P_{X\geq T}P_{Y=1|X\geq T}P_{t_x<X<T}P_{Y=1|X<T}\end{aligned}$$

Expand left hand side

$$\begin{aligned}&(P_{X>t}P_{Y=1|X>t})^2 - P_{X>t_x}P_{X>t}(P_{Y=1|X>t})^2 + (P_{t_x<X<t}P_{Y=1|X<t}))^2 - P_{X>t_x}(P_{t_x<X<t})(P_{Y=1|X<t})^2\\ <& -2P_{X\geq T}P_{Y=1|X\geq T}P_{t_x<X<T}P_{Y=1|X<T}\end{aligned}$$

Subtract $(P_{X\geq T}P_{Y=1|X\geq T})^2$ and $(P_{t_x<X<T}P_{Y=1|X<T}))^2$ from both sides

$$\begin{aligned}&-P_{X\geq t_x}P_{X\geq T}(P_{Y=1|X\geq T})^2 - P_{X\geq t_x}(P_{t_x<X<T})(P_{Y=1|X<T})^2\\ <& -(P_{X\geq T}P_{Y=1|X\geq T})^2 - (P_{t_x<X<T}P_{Y=1|X<T}))^2 - 2P_{X\geq T}P_{Y=1|X\geq T}P_{t_x<X<T}P_{Y=1|X<T}\end{aligned}$$

Factor right hand side

$$\begin{aligned}&-P_{X\geq t_x}P_{X\geq T}(P_{Y=1|X\geq T})^2 - P_{X\geq t_x}(P_{t_x<X<T})(P_{Y=1|X<T})^2\\ <& -((P_{X\geq T}P_{Y=1|X\geq T} + (P_{t_x<X<T})P_{Y=1|X<t}))^2\end{aligned}$$

Add $P_{X\geq t_x}(P_{t_x<X<T})P_{Y=1|X<T}$ and $P_{X\geq t_x}P_{X\geq T}P_{Y=1|X\geq T}$

$$\begin{aligned}&P_{X\geq t_x}P_{X\geq T}P_{Y=1|X} - P_{X\geq t_x}P_{X\geq T}(P_{Y=1|X\geq T})^2 + P_{X\geq t_x}(P_{t_x<X<T})P_{Y=1|x<T} - P_{X\geq t_x}(P_{t_x<X<T})(P_{Y=1|X<T})^2\\ <& P_{X\geq t_x}P_{X\geq T}P_{Y=1|X\geq T} + P_{X\geq t_x}(P_{t_x<X<T})P_{Y=1|X<T} - ((P_{X\geq t}P_{Y=1|X\geq T} + (P_{t_x<X<T})P_{Y=1|X<T}))^2\end{aligned}$$

Factor

$$P_{X\geq t_x}P_{X\geq T}P_{Y=1|X\geq T}(1-P_{Y=1|X\geq T})+P_{X\geq t_x}(P_{t_x<X<T})P_{Y=1|x<T}(1-P_{Y=1|X<T})$$
$$<(P_{X\geq T}P_{Y=1|X\geq T}+(P_{t_x<X<T})P_{Y=1|X<T})(P_{X\geq t_x}-(P_{X\geq T}P_{Y=1|X\geq T}+(P_{t_x<X<T})P_{Y=1|X<T}))$$

Divide by $P_{X\geq t_x}$

$$P_{X\geq T}P_{Y=1|X\geq T}(1-P_{Y=1|X\geq T})+(P_{t_x<X<T})P_{Y=1|x<T}(1-P_{Y=1|X<T})$$
$$<\frac{(P_{X\geq T}P_{Y=1|X\geq T}+(P_{t_x<x<T})P_{Y=1|X<T})(P_{X\geq t_x}-(P_{X\geq T}P_{Y=1|X\geq T}+(P_{t_x<x<T})P_{Y=1|X<T}))}{P_{X\geq t_x}}$$

Note $P_{t_x<X<T}=P_{X<T}-P_{X<t_x}$.

$$P_{X\geq T}P_{Y=1|X\geq T}(1-P_{Y=1|X\geq T})+(P_{X<T}-P_{X<t_x})P_{Y=1|x<T}(1-P_{Y=1|X<T})$$
$$<\frac{(P_{X\geq T}P_{Y=1|X\geq T}+(P_{t_x<x<T})P_{Y=1|X<T})(P_{X\geq t_x}-(P_{X\geq T}P_{Y=1|X\geq T}+(P_{t_x<x<T})P_{Y=1|X<T}))}{P_{X\geq t_x}}$$

Distribute

$$P_{X\geq T}P_{Y=1|X\geq T}(1-P_{Y=1|X\geq T})+P_{X<T}P_{Y=1|X<T}(1-P_{Y=1|X<T})-(P_{X<t_x}P_{Y=1|X<T})(1-P_{Y=1|X<T}))$$
$$<\frac{(P_{X\geq T}P_{Y=1|X\geq T}+(P_{t_x<X<T})P_{Y=1|X<T})(P_{X\geq t_x}-(P_{X\geq T}P_{Y=1|X\geq T}+(P_{t_x<X<T})P_{Y=1|X<T}))}{P_{X\geq t_x}}$$

Add $(P_{X<t_x}P_{Y=1|X<T})(1-P_{Y=1|X<T}))$ to both sides

$$P_{X\geq T}P_{Y=1|X\geq T}(1-P_{Y=1|X\geq T})+P_{X<T}P_{Y=1|X<T}(1-P_{Y=1|X<T})$$
$$<(\frac{(P_{X\geq T}P_{Y=1|X\geq T}+(P_{t_x<X<T})P_{Y=1|X<T})(P_{X\geq t_x}-(P_{X\geq T}P_{Y=1|X\geq T}+(P_{t_x<X<T})P_{Y=1|X<T}))}{P_{X\geq t_x}}$$
$$+(P_{X<t_x}P_{Y=1|X<T})(1-P_{Y=1|X<T}))$$

$$
\begin{aligned}
&(\frac{(P_{X\geq T}P_{Y=1|X\geq T})(P_{X\geq T}-P_{X\geq T}P_{Y=1|X\geq T})}{P_{X\geq T}}+\frac{(P_{X<T}P_{Y=1|X<T})(P_{X<T}-P_{X<T}P_{Y=1|X<T})}{P_{X<T}})\\
<&(\frac{(P_{X\geq T}P_{Y=1|X\geq T}+(P_{t_x<x<T})P_{Y=1|X<T})(P_{X\geq t_x}-(P_{X\geq T}P_{Y=1|X\geq T}+(P_{t_x<x<T})P_{Y=1|X<T}))}{P_{X\geq t_x}}\\
&+\frac{((P_{X<t_x})P_{Y=1|X<T})((P_{X<t_x})-(P_{X<t_x})P_{Y=1|X<T})}{(P_{X<t_x})})
\end{aligned}
$$

$$
\begin{aligned}
&-(\frac{(P_{X\geq T}P_{Y=1|X\geq T})(P_{X\geq T}-P_{X\geq T}P_{Y=1|X\geq T})}{P_{X\geq T}P_{Y=1|X\geq T}+P_{X\geq T}-P_{X\geq T}P_{Y=1|X\geq T}}+\frac{(P_{X<T}P_{Y=1|X<T})(P_{X<T}-P_{X<T}P_{Y=1|X<T})}{P_{X<T}P_{Y=1|X<T}+P_{X<T}-P_{X<T}P_{Y=1|X<T}})\\
>&-(\frac{(P_{X\geq T}P_{Y=1|X\geq T}+(P_{t_x<x<T})P_{Y=1|X<T})(P_{X\geq t_x}-(P_{X\geq T}P_{Y=1|X\geq T}-(P_{t_x<x<T})P_{Y=1|X<T}))}{P_{X\geq T}P_{Y=1|X\geq T}+(P_{t_x<x<T})P_{Y=1|X<T}+P_{X\geq t_x}-(P_{X\geq T}P_{Y=1|X\geq T}-(P_{t_x<x<T})P_{Y=1|X<T})}\\
&+\frac{((P_{X<t_x})P_{Y=1|X<T})((P_{X<t_x})-(P_{X<t_x})P_{Y=1|X<T})}{(P_{X<t_x})P_{Y=1|X<T}+(P_{X<t_x})-(P_{X<t_x})P_{Y=1|X<T}})
\end{aligned}
$$

$$(P_y(1-P_y))-(\frac{ab}{a+b}+\frac{cd}{c+d})>(P_y(1-P_y))-(\frac{ab}{a+b}+\frac{cd}{c+d})$$

Thus,

$$Gini_{t_x=t}>Gini_{t_x<t}$$

A.3.b Now consider the case where $P_{X>t_x}<P_{X>T}$. Start with what is given

$$P_{Y=1|X\geq T}>P_{Y=1|X<T}$$

Subtract $P_{Y=1|X<T}$ from both sides

$$0<P_{Y=1|X\geq T}-P_{Y=1|X<T}$$

Square both sides

$$0 < (P_{Y=1|X\geq T} - P_{Y=1|X<T})^2$$

Expand

$$0 < (P_{Y=1|X\geq T})^2 - 2P_{Y=1|X<T}P_{Y=1|X\geq T} + (P_{Y=1|X<T})^2$$

Add $2P_{Y=1|X<T}P_{Y=1|X\geq T}$ to both sides

$$2P_{Y=1|X<T}P_{Y=1|X\geq T} < (P_{Y=1|X\geq T})^2 + (P_{Y=1|X<T})^2$$

Multiply both sides by $P_{X<T}P_{t<X<t_x}($

$$2P_{X<T}P_{Y=1|X<T}P_{t<X<t_x}P_{Y=1|X\geq T} < P_{X<T}P_{t<X<t_x}(P_{Y=1|X\geq T})^2 + P_{t<X<t_x}P_{X<T}(P_{Y=1|X<T})^2$$

Multiply by -1

$$-2P_{X<T}P_{Y=1|X<T}P_{t<X<t_x}P_{Y=1|X\geq T} < (-P_{X<T})P_{t<X<t_x}(P_{Y=1|X\geq T})^2 + (-P_{t<X<t_x})P_{X<T}(P_{Y=1|X<T})^2$$

Note $-P_{t<X<t_x}$ is $(P_{t<X<t_x} - P_{X<t_x})$ to $-P_{X<T}$ and $(P_{X<T} - P_{X<t_x})$

$$\begin{aligned} &- 2P_{X<T}P_{Y=1|X<T}P_{t<X<t_x}P_{Y=1|X\geq T} \\ &< (P_{t<X<t_x} - P_{X<t_x})P_{t<X<t_x}(P_{Y=1|X\geq T})^2 + (P_{X<T} - P_{X<t_x})P_{X<T}(P_{Y=1|X<T})^2 \end{aligned}$$

Distribute

$$- 2P_{X<Tt}P_{Y=1|X<T}P_{t<X<t_x}P_{Y=1|X\geq T}$$
$$< -P_{X<t_x}P_{t<X<t_x}(P_{Y=1|X\geq T})^2 + P^2_{t<X<t_x}P^2_{Y=1|X\geq T} - P_{X<t_x}P_{X<T}(P_{Y=1|X<T})^2 + P^2_{X<T}P^2_{Y=1|X<T}$$

Subtract $P^2_{X<T}P^2_{Y=1|X<T}$ and $P^2_{t<X<t_x}P^2_{Y=1|X\geq T}$ from both sides

$$- P^2_{X<T}P^2_{Y=1|X<T} - P^2_{t<X<t_x}P^2_{Y=1|X\geq T} - 2P_{X<T}P_{Y=1|X<T}P_{t<X<t_x}P_{Y=1|X\geq T}$$
$$< -P_{X<t_x}P_{t<X<t_x}(P_{Y=1|X\geq T})^2 - P_{X<t_x}P_{X<T}(P_{Y=1|X<T})^2$$

Factor left hand side

$$- (P_{X<T}P_{Y=1|X<T} + (P_{t<X<t_x})P_{Y=1|X\geq T})^2$$
$$< -P_{X<t_x}P_{t<X<t_x}(P_{Y=1|X\geq T})^2 - P_{X<t_x}P_{X<t}(P_{Y=1|X<T})^2$$

Add $P_{X<t_x}P_{t<X<t_x}P_{Y=1|X\geq T}$ and $P_{X<t_x}P_{X<T}P_{Y=1|X<T}$ to both sides

$$(P_{X<T}P_{Y=1|X<T} + (P_{t<X<t_x})P_{Y=1|X\geq T})(P_{X<t_x} - (P_{X<T}P_{Y=1|X<T} - P_{t<X<t_x}P_{Y=1|X\geq T}))$$
$$< P_{X<t_x}P_{t<X<t_x}P_{Y=1|X\geq T}(1 - P_{Y=1|X\geq T}) + P_{X<t_x}P_{X<T}P_{Y=1|X<T}(1 - P_{Y=1|X<T})$$

Divide by $P_{X<t_x}$

$$\frac{(P_{X<T}P_{Y=1|X<T} + (P_{t<X<t_x})P_{Y=1|X\geq T})(P_{X<t_x} - (P_{X<T}P_{Y=1|X<T} - P_{t<X<t_x}P_{Y=1|X\geq T}))}{P_{X<t_x}}$$
$$< P_{t<X<t_x}P_{Y=1|X\geq T}(1 - P_{Y=1|X\geq T}) + P_{X<T}P_{Y=1|X<T}(1 - P_{Y=1|X<T})$$

Separate $P_{t<X<t_x}$ term

$$\frac{(P_{X<t}P_{Y=1|X<t} + (P_{t<X<t_x})P_{Y=1|X>t})(P_{X<t_x} - (P_{X<t}P_{Y=1|X<t} - P_{t<X<t_x}P_{Y=1|X>t}))}{P_{X<t_x}}$$

$$< P_{X\geq T}P_{Y=1|X\geq T}(1 - P_{Y=1|X\geq T}) + P_{X<T}P_{Y=1|X<T}(1 - P_{Y=1|X<T}) - P_{X>t_x}P_{Y=1|X\geq T}(1 - P_{Y=1|X\geq T})$$

Add $P_{X>t_x}P_{Y=1|X\geq T}(1 - P_{Y=1|X\geq T})$ to both sides

$$P_{X>t_x}P_{Y=1|X\geq T}(1 - P_{Y=1|X\geq T})$$
$$+ \frac{(P_{X<T}P_{Y=1|X<T} + (P_{t<X<t_x})P_{Y=1|X\geq T})(P_{X<t_x} - (P_{X<T}P_{Y=1|X<T} - P_{t<X<t_x}P_{Y=1|X\geq T}))}{P_{X<t_x}}$$
$$< P_{X\geq T}P_{Y=1|X\geq T}(1 - P_{Y=1|X\geq T}) + P_{X<T}P_{Y=1|X<T}(1 - P_{Y=1|X<T})$$

Multiply by $P_{X>t_x}$ and $P_{X\geq T}$. Divide by $P_{X>t_x}P_{Y=1|X\geq T} + (P_{X>t_x} - P_{X>t_x}P_{Y=1|X\geq T}$ and $P_{X<t}P_{Y=1|X<T} + (P_{t<X<t_x})P_{Y=1|X\geq T} + (P_{X<t_x} - (P_{X<T}P_{Y=1|X<T} - P_{t<X<t_x}P_{Y=1|X>t}))$

$$\frac{P_{X>t_x}P_{Y=1|X\geq T}(P_{X>t_x} - P_{X>t_x}P_{Y=1|X>t})}{P_{X>t_x}P_{Y=1|X\geq T} + (P_{X>t_x} - P_{X>t_x}P_{Y=1|X\geq T})}$$
$$+ \frac{(P_{X<t}P_{Y=1|X<T} + (P_{t<X<t_x})P_{Y=1|X\geq T})(P_{X<t_x} - (P_{X<T}P_{Y=1|X<T} - P_{t<X<t_x}P_{Y=1|X\geq T}))}{P_{X<T}P_{Y=1|X<T} + (P_{t<X<t_x})P_{Y=1|X\geq T} + (P_{X<t_x} - (P_{X<T}P_{Y=1|X<T} - P_{t<X<t_x}P_{Y=1|X\geq T}))}$$
$$< \frac{(P_{X\geq T}P_{Y=1|X\geq T})(P_{X\geq T} - P_{X\geq T}P_{Y=1|X\geq T})}{P_{X\geq T}P_{Y=1|X\geq T} + P_{X\geq T} - P_{X\geq T}P_{Y=1|X\geq T}} + \frac{(P_{X<T}P_{Y=1|X<T})(P_{X<T} - P_{X<T}P_{Y=1|X<T})}{P_{X<T}P_{Y=1|X<T} + P_{X<T} - P_{X<T}P_{Y=1|X<T}}$$

Thus $Gini_{t_x>T} < Gini_{t_x=T}$ If the expression for $Gini_{t_x=T}$ is greater than the expression for $Gini_{t_x<T}$ and the expression for $Gini_{t_x=T}$ is greater than the expression for $Gini_{t_x>T}$ then it shows that the Gini Index is the highest when $t_x = T$.

A.4 *Proof of Theorem 1 for the chi-square Statistic*

Let X be a random variable and Y a dichotomous variable. Also, let T be a threshold such that, $P_{Y=1|X\geq T} > P_{Y=1|X<T}$. There are three possibe cases that can occur when selecting a threshold for X, t_x: (1) $t_x < T$,(2) $t_x = T$, and (3) $t_x > T$. The expression for the chi-square,

$\frac{(ad-bc)^2}{(a+b)(c+d)(b+d)(a+c)}$, for each case can be found using the expressions for a,b,c, and defined in equations 6.1, 6.2, and 6.3. We can then show that the chi-square is maximized when $t_x = T$.

A.4.a Consider the case where $P_{X>t_x} > P_{X>T}$. Start with what is given

$$P_{Y=1|X\geq T} > P_{Y=1|X<T}$$

Subtract $P_{Y=1|X<T}$ from both sides

$$P_{Y=1|X\geq T} - P_{Y=1|X<T} > 0$$

Square both sides

$$(P_{Y=1|X\geq T} - P_{Y=1|X<T})^2 > 0$$

Square both sides

$$(P_{Y=1|X\geq T})^2 - 2P_{Y=1|X\geq T}P_{Y=1|X<T} + (P_{Y=1|X<T})^2 > 0$$

Add $2P_{Y=1|X\geq T}P_{Y=1|X<T}$ to both sides

$$(P_{Y=1|X\geq T})^2 + (P_{Y=1|X<T})^2 > 2P_{Y=1|X\geq T}P_{Y=1|X<T}$$

Multiply both sides by $(P_{x<T})^2 - (P_{x<t_x})^2$

$$((P_{x<T})^2 - (P_{x<t_x})^2)(P_{Y=1|X\geq T})^2 + ((P_{x<T})^2 - (P_{x<t_x})^2)(P_{Y=1|X<T})^2$$
$$> 2((P_{x<T})^2 - (P_{x<t_x})^2)P_{Y=1|X\geq T}P_{Y=1|X<T}$$

Distribute

$$(P_{x<T})^2(P_{Y=1|X\geq T})^2 - (P_{x<t_x})^2(P_{Y=1|X\geq T})^2 + (P_{x<T})^2(P_{Y=1|X<T})^2 - (P_{x<t_x})^2(P_{Y=1|X<T})^2$$
$$> 2(P_{x<T})^2P_{Y=1|X\geq T}P_{Y=1|X<T} - 2(P_{x<t_x})^2P_{Y=1|X\geq T}P_{Y=1|X<T}$$

Subtract $2(P_{x<T})^2 P_{Y=1|X\geq T} P_{Y=1|X<T}$, add $(P_{x<t_x})^2 (P_{Y=1|X\geq T})^2$ and $(P_{x<t_x})^2 (P_{Y=1|X<T})^2$.

$$(P_{x<T})^2 (P_{Y=1|X\geq T})^2 + (P_{x<T})^2 (P_{Y=1|X<T})^2 - 2(P_{x<T})^2 P_{Y=1|X\geq T} P_{Y=1|X<T}$$
$$> (P_{x<t_x})^2 (P_{Y=1|X\geq T})^2 + (P_{x<t_x})^2 (P_{Y=1|X<T})^2 - 2(P_{x<t_x})^2 P_{Y=1|X\geq T} P_{Y=1|X<T}$$

Multiply both sides by $(P_{X\geq T})^2$.

$$(P_{X\geq T})^2 (P_{x<T})^2 (P_{Y=1|X\geq T})^2 + (P_{X\geq T})^2 (P_{x<T})^2 (P_{Y=1|X<T})^2 - 2(P_{X\geq T})^2 (P_{x<T})^2 P_{Y=1|X\geq T} P_{Y=1|X<T}$$
$$> (P_{X\geq T})^2 (P_{x<t_x})^2 (P_{Y=1|X\geq T})^2 + (P_{X\geq T})^2 (P_{x<t_x})^2 (P_{Y=1|X<T})^2 - 2(P_{X\geq T})^2 (P_{x<t_x})^2 P_{Y=1|X\geq T} P_{Y=1|X<T}$$

Factor by difference of squares

$$(P_{X<T} P_{X\geq T} P_{Y=1|X\geq T} - P_{X<T} P_{X\geq T} P_{Y=1|X<T})^2 > (P_{X<t_x} P_{X\geq T} P_{Y=1|X\geq T} - P_{X<t_x} P_{X\geq T} P_{Y=1|X<T})^2$$

On the left side, add and subtract $P_{X\geq T} P_{Y=1|X\geq T} P_{X<T} P_{Y=1|X<T}$

$$(P_{X<T} P_{X\geq T} P_{Y=1|X\geq T} - P_{X\geq T} P_{Y=1|X\geq T} P_{X<T} P_{Y=1|X<T}$$
$$- P_{X<T} P_{X\geq T} P_{Y=1|X<T} + P_{X\geq T} P_{Y=1|X\geq T} P_{X<T} P_{Y=1|X<T})^2$$
$$> (P_{X<t_x} P_{X\geq T} P_{Y=1|X\geq T} - P_{X<t_x} P_{X\geq T} P_{Y=1|X<T})^2$$

Thus, the left side factors by difference of squares

$$(P_{X\geq T} P_{Y=1|X\geq T} (P_{X<T} - P_{X<T} P_{Y=1|X<T}) - (P_{X\geq T} - P_{X\geq T} P_{Y=1|X\geq T})(P_{X<T} P_{Y=1|X<T}))^2$$
$$> (P_{X<t_x} P_{X\geq T} P_{Y=1|X\geq T} - P_{X<t_x} P_{X\geq T} P_{Y=1|X<T})^2$$

On the right side, note $-P_{X\geq T} = (P_{X<T} - 1)$

$$(P_{X\geq T}P_{Y=1|X\geq T}(P_{X<T} - P_{X<T}P_{Y=1|X<T}) - (P_{X\geq T} - P_{X\geq T}P_{Y=1|X\geq T})(P_{X<T}P_{Y=1|X<T}))^2$$
$$> (P_{X<t_x}P_{X\geq T}P_{Y=1|X\geq T} + P_{X<t_x}P_{Y=1|X<T}(P_{X<T} - 1))^2$$

Distribute on the right

$$(P_{X\geq T}P_{Y=1|X\geq T}(P_{X<T} - P_{X<T}P_{Y=1|X<T}) - (P_{X\geq T} - P_{X\geq T}P_{Y=1|X\geq T})(P_{X<T}P_{Y=1|X<T}))^2$$
$$> (P_{X<t_x}P_{X\geq T}P_{Y=1|X\geq T} + P_{X<t_x}P_{Y=1|X<T}P_{X<T} - P_{X<t_x}P_{Y=1|X<T})^2$$

Also note $P_{X<t_x} + P_{X>t_x} = 1$. So, multiply by 1 on the right

$$(P_{X\geq T}P_{Y=1|X\geq T}(P_{X<T} - P_{X<T}P_{Y=1|X<T}) - (P_{X\geq T} - P_{X\geq T}P_{Y=1|X\geq T})(P_{X<T}P_{Y=1|X<T}))^2$$
$$> (P_{X<t_x}P_{X\geq T}P_{Y=1|X\geq T} + P_{X<t_x}P_{Y=1|X<T}P_{X<T} - P_{X<t_x}P_{Y=1|X<T}(P_{X<t_x} + P_{X>t_x}))^2$$

Distribute on the right

$$(P_{X\geq T}P_{Y=1|X\geq T}(P_{X<T} - P_{X<T}P_{Y=1|X<T}) - (P_{X\geq T} - P_{X\geq T}P_{Y=1|X\geq T})(P_{X<T}P_{Y=1|X<T}))^2$$
$$> (P_{X<t_x}P_{X\geq T}P_{Y=1|X\geq T} + P_{X<t_x}P_{Y=1|X<T}P_{X<T}$$
$$- P_{X<t_x}P_{Y=1|X<T}P_{X<t_x} - P_{X<t_x}P_{Y=1|X<T}P_{X>t_x})^2$$

On the right side, add and subtract $(P_{X\geq T}P_{Y=1|X\geq T}+(P_{X<T}-P_{X<t_x})P_{Y=1|X<T})P_{X<t_x}P_{Y=1|X<T}$.

$$(P_{X\geq T}P_{Y=1|X\geq T}(P_{X<T} - P_{X<T}P_{Y=1|X<T}) - (P_{X\geq T} - P_{X\geq T}P_{Y=1|X\geq T})(P_{X<T}P_{Y=1|X<T}))^2$$
$$> (P_{X<t_x}P_{X\geq T}P_{Y=1|X\geq T} + P_{X<t_x}P_{Y=1|X<T}P_{X<T}$$
$$- (P_{X\geq T}P_{Y=1|X\geq T} + (P_{X<T} - P_{X<t_x})P_{Y=1|X<T})P_{X<t_x}P_{Y=1|X<T}$$
$$- P_{X<t_x}P_{Y=1|X<T}P_{X<t_x} - P_{X<t_x}P_{Y=1|X<T}P_{X\geq t_x}$$
$$+ (P_{X\geq T}P_{Y=1|X\geq T} + (P_{X<T} - P_{X<t_x})P_{Y=1|X<T})P_{X<t_x}P_{Y=1|X<T})^2$$

Factor the right side,

$$(P_{X\geq T}P_{Y=1|X\geq T}(P_{X<T} - P_{X<T}P_{Y=1|X<T}) - (P_{X\geq T} - P_{X\geq T}P_{Y=1|X\geq T})(P_{X<T}P_{Y=1|X<T}))^2$$
$$> ((P_{X\geq T}P_{Y=1|X\geq T} + (P_{X<T} - P_{X<T})P_{Y=1|X<T})((P_{X<T} - P_{X<T})P_{Y=1|X<T})$$
$$- (P_{X\geq t_x} - (P_{X\geq T}P_{Y=1|X\geq T} + (P_{X<T} - P_{X<T})P_{Y=1|X<T}))(P_{X<t_x}P_{Y=1|X<T}))^2$$

Divide both sides by $P_{X\geq T}(1 - P_{Y=1})P_{X<T}P_{Y=1}$ and we have

$$\chi^2_{t_x=T} > \chi^2_{t_x<T}$$

A.4.b Now consider the case where $P_{X>t_x} < P_{X>T}$. Start with what is given

$$P_{Y=1|X\geq T} > P_{Y=1|X<T}$$

Subtract $P_{Y=1|X<T}$ from both sides

$$P_{Y=1|X\geq T} - P_{Y=1|X<T} > 0$$

Square both sides

$$(P_{Y=1|X\geq T} - P_{Y=1|X<T})^2 > 0$$

Square both sides

$$(P_{Y=1|X\geq T})^2 - 2P_{Y=1|X\geq T}P_{Y=1|X<T} + (P_{Y=1|X<T})^2 > 0$$

Add $2P_{Y=1|X\geq T}P_{Y=1|X<T}$ to both sides

$$(P_{Y=1|X\geq T})^2 + (P_{Y=1|X<T})^2 > 2P_{Y=1|X\geq T}P_{Y=1|X<T}$$

Multiply both sides by $(P_{x>T})^2 - (P_{x>t_x})^2$

$$\begin{aligned}&((P_{x>T})^2 - (P_{x>t_x})^2)(P_{Y=1|X\geq T})^2 + ((P_{x>T})^2 - (P_{x>t_x})^2)(P_{Y=1|X<T})^2\\ &> 2((P_{x>T})^2 - (P_{x>t_x})^2)P_{Y=1|X\geq T}P_{Y=1|X<T}\end{aligned}$$

Distribute

$$\begin{aligned}&(P_{x>T})^2(P_{Y=1|X\geq T})^2 - (P_{x>t_x})^2(P_{Y=1|X\geq T})^2 + (P_{x>T})^2(P_{Y=1|X<T})^2 - (P_{x>t_x})^2(P_{Y=1|X<T})^2\\ &> 2(P_{x>T})^2P_{Y=1|X\geq T}P_{Y=1|X<T} - (P_{x>t_x})^2P_{Y=1|X\geq T}P_{Y=1|X<T}\end{aligned}$$

Rearrange terms

$$\begin{aligned}&(P_{x>T})^2(P_{Y=1|X\geq T})^2 - 2(P_{x>T})^2P_{Y=1|X\geq T}P_{Y=1|X<T} + (P_{x>T})^2(P_{Y=1|X<T})^2\\ &> (P_{x>t_x})^2(P_{Y=1|X\geq T})^2 - 2(P_{x>t_x})^2P_{Y=1|X\geq T}P_{Y=1|X<T} + (P_{x>t_x})^2(P_{Y=1|X<T})^2\end{aligned}$$

Multiply by $(P_{X<T})^2$

$$\begin{aligned}&(P_{x>T})^2(P_{X<T})^2(P_{Y=1|X\geq T})^2 - 2(P_{X<T})^2(P_{x>T})^2P_{Y=1|X\geq T}P_{Y=1|X<T} + (P_{x>T})^2(P_{X<T})^2(P_{Y=1|X<T})^2\\ &> (P_{x>t_x})^2(P_{X<T})^2(P_{Y=1|X\geq T})^2 - 2(P_{X<T})^2(P_{x>t_x})^2P_{Y=1|X\geq T}P_{Y=1|X<T} + (P_{X<T})^2(P_{x>t_x})^2(P_{Y=1|X<T})^2\end{aligned}$$

Factor

$$((P_{x>T})(P_{X<T})(P_{Y=1|X\geq T}) - (P_{x>T})^2(P_{X<T})(P_{Y=1|X<T}))^2$$
$$> ((P_{x>t_x})(P_{X<T})(P_{Y=1|X\geq T}) - (P_{X<T})(P_{x>t_x})(P_{Y=1|X<T}))^2$$

Therefore,

$$(a_{t_{x=T}}d_{t_{x=T}} - b_{t_{x=T}}c_{t_{x=T}})^2 > (a_{t_{x>T}}d_{t_{x>T}} - b_{t_{x>T}}c_{t_{x>T}})^2$$

and we have, $\chi^2_{t_x=T} > \chi^2_{t_x>T}$ If the expression for $chi^2_{t_x=T}$ is greater than the expression for $chi^2_{t_x<T}$ and the expression for $chi^2_{t_x=T}$ is greater than the expression for $chi^2_{t_x>T}$ then it shows that the chi-square is the highest when $t_x = T$.

A.5 *Proof of Theorem 1 for Relative Risk*

Let X be a random variable and Y a dichotomous variable. Also, let T be a threshold such that, $P_{Y=1|X\geq T} > P_{Y=1|X<T}$. There are three possibe cases that can occur when selecting a threshold for X, t_x: (1) $t_x < T$,(2) $t_x = T$, and (3) $t_x > T$. The expression for the Relative Risk, $\frac{a/(a+b)}{c/(c+d)}$, for each case can be found using the expressions for a,b,c, and defined in equations 6.1, 6.2, and 6.3. We can then show that the Relative Risk is maximized when $t_x = T$.

A.5.a Consider the case where $P_{X>t_x} > P_{X>T}$. Start with what is given

$$P_{Y=1|X\geq T} > P_{Y=1|X<T}$$

Multiply both sides by $P_{Y=1|X<T}$

$$P_{Y=1|X\geq T}P_{Y=1|X<T} > (P_{Y=1|X<T})^2$$

Set equal to 0

$$0 > -P_{Y=1|X\geq T}P_{Y=1|X<T} + (P_{Y=1|X<T})^2$$

Multiply both sides by $(P_{X<t_x} - P_{X<T})$

$$0 > -P_{Y=1|X\geq T}P_{Y=1|X<T}(P_{X<t_x} - P_{X<T}) + (P_{X<t_x} - P_{X<T})(P_{Y=1|X<T})^2$$

Replace $P_{X<t_x}$ with $(1 - P_{X>t_x})$ and $P_{X<T}$ with $(1 - P_{X\geq T})$

$$0 > -P_{Y=1|X\geq T}P_{Y=1|X<T}((1 - P_{X>t_x}) - (1 - P_{X\geq T})) + (P_{X<t_x} - P_{X<T})(P_{Y=1|X<T})^2$$

Distribute

$$0 > -P_{Y=1|X\geq T}P_{Y=1|X<T}P_{X>t_x}+P_{Y=1|X\geq T}P_{Y=1|X<T}P_{X\geq T})+(P_{X<t_x}-P_{X<T})(P_{Y=1|X<T})^2$$

Add $P_{Y=1|X\geq T}P_{Y=1|X<T}P_{X>t_x}$ to both sides

$$P_{Y=1|X\geq T}P_{Y=1|X<T}P_{X>t_x} > P_{Y=1|X\geq T}P_{Y=1|X<T}P_{X\geq T}) + (P_{X<t_x} - P_{X<T})(P_{Y=1|X<T})^2$$

Factor out $P_{Y=1|X<T}$ from the left side

$$P_{Y=1|X\geq T}P_{Y=1|X<T}P_{X>t_x} > P_{Y=1|X<T}(P_{Y=1|X\geq T}P_{X\geq T} + (P_{X<t_x} - P_{X<T})P_{Y=1|X<T})$$

Divide both sides by $(P_{Y=1|X<T})^2P_{X>t_x}$

$$\frac{P_{Y=1|X\geq T}}{P_{Y=1|X<T}} > \frac{P_{Y=1|X\geq T}P_{X\geq T} + (P_{X<t_x} - P_{X<T})P_{Y=1|X<T}}{P_{Y=1|X<T}P_{X>t_x}}$$

Multiply the left side by $\frac{P_{X\geq T}P_{X<T}}{P_{X\geq T}P_{X<T}}$ and the right side by $\frac{P_{X<t_x}}{P_{X<t_x}}$

$$\frac{P_{X\geq T}P_{Y=1|X\geq T}P_{X<T}}{P_{X<T}P_{Y=1|X<T}P_{X\geq T}} > \frac{P_{Y=1|X\geq T}P_{X\geq T} + (P_{X<t_x} - P_{X<T})P_{Y=1|X<T}P_{X<t_x}}{P_{X<t_x}P_{Y=1|X<T}P_{X>t_x}}$$

Thus $RR_{t_x=T} > RR_{t_x<T}$

A.5.b Now consider the case where $P_{X>t_x} < P_{X>T}$. Start with what is given

$$P_{Y=1|X\geq T} > P_{Y=1|X<T}$$

Multiply both sides by $P_{Y=1|X\geq T}$

$$(P_{Y=1|X\geq T})^2 > P_{Y=1|X\geq T}P_{Y=1|X<T}$$

Set equal to 0

$$(P_{Y=1|X\geq T})^2 - P_{Y=1|X\geq T}P_{Y=1|X<T} > 0$$

Multiply both sides by $(P_{X<t_x} - P_{X<T})$

$$(P_{X<t_x} - P_{X<T})(P_{Y=1|X\geq T})^2 - (P_{X<t_x} - P_{X<T})P_{Y=1|X\geq T}P_{Y=1|X<T} > 0$$

Factor out a negative 1

$$(P_{X<t_x} - P_{X<T})(P_{Y=1|X\geq T})^2 + (P_{X<T} - P_{X<t_x})P_{Y=1|X\geq T}P_{Y=1|X<T} > 0$$

Distribute

$$P_{X<t_x}(P_{Y=1|X\geq T})^2 - P_{X<T}(P_{Y=1|X\geq T})^2 + P_{X<T}P_{Y=1|X\geq T}P_{Y=1|X<T} - P_{X<t_x}P_{Y=1|X\geq T}P_{Y=1|X<T} > 0$$

Add $P_{X<t_x}P_{Y=1|X\geq T}P_{Y=1|X<T}$ to both sides

$$P_{X<t_x}(P_{Y=1|X\geq T})^2 - P_{X<T}(P_{Y=1|X\geq T})^2 + P_{X<T}P_{Y=1|X\geq T}P_{Y=1|X<T} > P_{X<t_x}P_{Y=1|X\geq T}P_{Y=1|X<T}$$

Factor out $P_{Y=1|X\geq T}$ from the left

$$P_{Y=1|X\geq T}(P_{X<t_x}(P_{Y=1|X\geq T}) - P_{X<T}(P_{Y=1|X\geq T}) + P_{X<T}P_{Y=1|X<T}) > P_{X<t_x}P_{Y=1|X\geq T}P_{Y=1|X<T}$$

Divide both sides by $P_{Y=1|X<T}$ and $(P_{X<t_x}(P_{Y=1|X\geq T}) - P_{X<T}(P_{Y=1|X\geq T}) + P_{X<T}P_{Y=1|X<T})$

$$\frac{P_{Y=1|X\geq T}}{P_{Y=1|X<T}} > \frac{P_{X<t_x}P_{Y=1|X\geq T}}{(P_{X<t_x}(P_{Y=1|X\geq T}) - P_{X<T}(P_{Y=1|X\geq T}) + P_{X<T}P_{Y=1|X<T})}$$

Multiply by $\frac{P_{X>T}}{P_{X>T}}$, $\frac{P_{X<T}}{P_{X<T}}$ and $\frac{P_{X>t_x}}{P_{X>t_x}}$

$$\frac{P_{X>T}P_{Y=1|X\geq T}P_{X<T}}{P_{X<T}P_{Y=1|X<T}P_{X>T}} > \frac{P_{X>t_x}P_{X<t_x}P_{Y=1|X\geq T}}{(P_{X<t_x}(P_{Y=1|X\geq T}) - P_{X<T}(P_{Y=1|X\geq T}) + P_{X<T}P_{Y=1|X<T})P_{X>t_x}}$$

Thus $RR_{t_x=T} > RR_{t_x>T}$ If the expression for $RR_{t_x=T}$ is greater than the expression for $RR_{t_x<T}$ and the expression for $RR_{t_x=T}$ is greater than the expression for $RR_{t_x>T}$ then it shows that the Relative Risk is the highest when $t_x = T$.

A.6 *Proof of theorem 1 for Kappa statistic*

Let X be a random variable and Y a dichotomous variable. Also, let T be a threshold such that, $P_{Y=1|X\geq T} > P_{Y=1|X<T}$. There are three possibe cases that can occur when selecting a threshold for X, t_x: (1) $t_x < T$,(2) $t_x = T$, and (3) $t_x > T$. The expression for Kappa, $\frac{(a+d)-((a+b)(a+c)+(c+d)(b+d))}{1-((a+b)(a+c)+(c+d)(b+d))}$, for each case can be found using the expressions for a,b,c, and defined in equations 6.1, 6.2, and 6.3. We can then show that Kappa is maximized when $t_x = T$.

A.6.a First we want to show that $Kappa_{t_x=T} > Kappa_{t_x<T}$

We begin with the true statement

$$P_{X<T} > P_{X<t_x}$$

Note $(P_{X\geq t_x} + P_{X<t_x}) = 1$

$$P_{X<T}(P_{X\geq t_x} + P_{X<t_x}) > P_{X<t_x}$$

Distribute and set equal to 0

$$P_{X<T}P_{X\geq t_x} - P_{X<t_x} + P_{X<T}P_{X<t_x} > 0$$

Factor out $P_{X<t_x}$

$$P_{X<T}P_{X\geq t_x} - P_{X<t_x}(1 - P_{X<T}) > 0$$

Note $1 - P_{X<T} = P_{X\geq T}$

$$P_{X<T}P_{X\geq t_x} - P_{X<t_x}(P_{X\geq T}) > 0$$

Add $P_{X<t_x}(P_{X\geq T})$ to both sides

$$P_{X<T}P_{X\geq t_x} > P_{X<t_x}(P_{X\geq T})$$

Multiply both sides by $(1 - P_{Y=1})$

$$P_{X<T}P_{X\geq t_x}(1 - P_{Y=1}) > P_{X<t_x}(P_{X\geq T})(1 - P_{Y=1})$$

Distribute

$$P_{X<T}P_{X\geq t_x} - P_{X<T}P_{X\geq t_x}P_{Y=1} > P_{X<t_x}P_{X\geq T} - P_{X<t_x}P_{X\geq T}P_{Y=1}$$

Add $P_{X<T}P_{Y=1}P_{X<t_x}$ to both sides

$$P_{X<T}P_{X\geq t_x}+P_{X<T}P_{Y=1}P_{X<t_x}-P_{X<T}P_{X\geq t_x}P_{Y=1} > P_{X<t_x}P_{X\geq T}+P_{X<t_x}P_{Y=1}P_{X<T}-P_{X<t_x}P_{X\geq T}P_{Y=1}$$

Factor both sides

$$P_{X<T}P_{X\geq t_x} + P_{X<T}P_{Y=1}(P_{X<t_x} - P_{X\geq t_x}) > P_{X<t_x}P_{X\geq T} + P_{X<t_x}P_{Y=1}(P_{X<T} - P_{X\geq T})$$

Factor $P_{X<T}$ and P_{t_x}

$$P_{X<T}(P_{X\geq t_x} + P_{Y=1}(P_{X<t_x} - P_{X\geq t_x})) > P_{X<t_x}(P_{X\geq T} + P_{Y=1}(P_{X<T} - P_{X\geq T}))$$

Divide by $P_{X\geq t_x} + P_{Y=1}(P_{X<t_x} - P_{X\geq t_x})$ and $P_{X\geq T} + P_{Y=1}(P_{X<T} - P_{X\geq T})$

$$\frac{P_{X<T}}{P_{X\geq T} + P_{Y=1}(P_{X<T} - P_{X\geq T})} > \frac{P_{X<t_x}}{P_{X\geq t_x} + P_{Y=1}(P_{X<t_x} - P_{X\geq t_x})}$$

Multiply by $2P_{X\geq T}(P_{Y=1|X\geq T} - P_{Y=1|X<T})$

$$\frac{2P_{X\geq T}P_{X<T}(P_{Y=1|X\geq T} - P_{Y=1|X<T})}{P_{X\geq T} + P_{Y=1}(P_{X<T} - P_{X\geq T})} > \frac{2P_{X<t_x}P_{X\geq T}(P_{Y=1|X\geq T} - P_{Y=1|X<T})}{P_{X\geq t_x} + P_{Y=1}(P_{X<t_x} - P_{X\geq t_x})}$$

Thus $Kappa_{t_x=T} > Kappa_{t_x<T}$. If the expression for $Kappa_{t_x=T}$ is greater than the expression for $Kappa_{t_x<T}$ and the expression for $Kappa_{t_x=T}$ is greater than the expression for $Kappa_{t_x>T}$ then it shows that the Kappa is the highest when $t_x = T$.

A.6.b

Next we show $Kappa_{t_x=T} < Kappa_{t_x<T}$

$$P_{X\geq T} > P_{X\geq t_x}$$

Note that $P_{X<t_x} + P_{X \geq t_x} = 1$

$$P_{X \geq T}(P_{X<t_x} + P_{X \geq t_x}) > P_{X \geq t_x}$$

Distribute and set equal to zero

$$P_{X \geq T} P_{X<t_x} - P_{X \geq t_x} + P_{X \geq T} P_{X \geq t_x} > 0$$

Factor out $P_{X \geq t_x}$

$$P_{X \geq T} P_{X<t_x} - (1 - P_{X \geq T}) P_{X \geq t_x} > 0$$

Note $1 - P_{X \geq T} = P_{X<T}$,

$$P_{X \geq T} P_{X<t_x} - P_{X<T} P_{X \geq t_x} > 0$$

Add $P_{X<T} P_{X \geq t_x}$ to both sides

$$P_{X \geq T} P_{X<t_x} > P_{X<T} P_{X \geq t_x}$$

Subtract $P_{X \geq T} P_{X \geq t_x}$ from both sides

$$P_{X \geq T} P_{X<t_x} - P_{X \geq T} P_{X \geq t_x} > P_{X<T} P_{X \geq t_x} - P_{X \geq T} P_{X \geq t_x}$$

Factor each side and multiply by 2

$$2P_{X \geq T}(P_{X<t_x} - P_{X \geq t_x}) > 2P_{X \geq t_x}(P_{X<t_x} - P_{X \geq t_x})$$

Multiply both sides by $(P = Y = 1 | X \geq T - P_{Y=1})$

$$2P_{X \geq T}(P = Y = 1 | X \geq T - P_{Y=1})(P_{X<t_x} - P_{X \geq t_x}) > 2P_{X \geq t_x}(P = Y = 1 | X \geq T - P_{Y=1})(P_{X<t_x} - P_{X \geq t_x})$$

Add $2P_{X\geq T}(P = Y = 1|X \geq T - P_{Y=1})P_{X\geq t_x}$ to both sides

$$2P_{X\geq T}(P = Y = 1|X \geq T - P_{Y=1})P_{X\geq t_x} + 2P_{X\geq T}(P = Y = 1|X \geq T - P_{Y=1})(P_{X<t_x} - P_{X\geq t_x}) >$$
$$2P_{X\geq T}(P = Y = 1|X \geq T - P_{Y=1})P_{X\geq t_x} + 2P_{X\geq t_x}(P = Y = 1|X \geq T - P_{Y=1})(P_{X<t_x} - P_{X\geq t_x})$$

Factor both sides

$$(2P_{X\geq T}(P_{Y=1|X\geq T} - P_{Y=1}))(P_{X\geq t_x} + P_{Y=1})(P_{X<t_x} - P_{X\geq t_x})) >$$
$$(2P_{X\geq t_x}(P_{Y=1|X\geq T} - P_{Y=1}))(P_{X\geq T} + P_{Y=1})(P_{X<T} - P_{X\geq T}))$$

Divide both sides by $P_{X\geq t_x} + P_{Y=1})(P_{X<t_x} - P_{X\geq t_x})$ and $P_{X\geq T} + P_{Y=1})(P_{X<T} - P_{X\geq T})$

$$\frac{2P_{X\geq T}(P_{Y=1|X\geq T} - P_{Y=1})}{P_{X\geq T} + P_{Y=1})(P_{X<T} - P_{X\geq T})} > \frac{2P_{X\geq t_x}(P_{Y=1|X\geq T} - P_{Y=1})}{P_{X\geq t_x} + P_{Y=1})(P_{X<t_x} - P_{X\geq t_x})}$$

Thus, $Kappa_{t_x=T} > Kappa_{t_x>T}$.

6.2 B: Proofs for Chapter 3

Theorem 3.1 *For continuous variables X_1 and X_2 and a dichotomous variable Y with prevalence $P_{Y=1}$ and thresholds T_1 and T_2 such that $P_{Y|T} > P_{Y|F}$ (Equation 3.1), then the inequality $g(t_1|X_1) < g(T_1|X_1)$ for all $t_1 \neq T_1$ holds where $g(T|X_1)$ is any of the six statistics defined in Table 3.3.*

B.1.1 *Relative Risk*

First we prove the theorem for $t_{x_1} < T$ and begin with the statement $P_{Y|T} > P_{Y|F}$. Multiplying both sides by $P_{t_{x_1}<X_1<T_1,X_2\geq T_2}P_{X_1\geq T_1,X_2\geq T_2}$ yields

$$P_{t_{x_1}<X_1<T_1,X_2\geq T_2}P_{X_1\geq T_1,X_2\geq T_2}P_{Y|T} > P_{t_{x_1}<X_1<T_1,X_2\geq T_2}P_{X_1\geq T_1,X_2\geq T_2}P_{Y|F}.$$

Replacing $P_{X_1\geq T_1,X_2\geq T_2}$ with $P_{X_1\geq T_1} - P_{X_1\geq T_1,X_2<T_2}$ and distributing yields

$$\begin{aligned}&P_{t_{x_1}<X_1<T_1,X_2\geq T_2}P_{X_1\geq T_1,X_2\geq T_2}P_{Y|T} > \\ &- P_{t_{x_1}<X_1<T_1,X_2\geq T_2}P_{X_1\geq T_1,X_2<T_2}P_{Y|F} + P_{X_1\geq T_1}P_{t_{x_1}<X_1<T_1,X_2\geq T_2}P_{Y|F}.\end{aligned}$$

Replacing $P_{t_{x_1}<X_1<T_1,X_2\geq T_2}$ with $P_{X_1\geq t_{X_1}} - P_{X_1\geq T_1}$ and distributing again yields,

$$\begin{aligned}&P_{X_1\geq t_{X_1}}P_{X_1\geq T_1,X_2\geq T_2}P_{Y|T} - P_{X_1\geq T_1}P_{X_1\geq T_1,X_2\geq T_2}P_{Y|T} > \\ &P_{X_1\geq T_1}P_{X_1\geq T_1,X_2<T_2}P_{Y|F} - P_{X_1} \geq t_{X_1}P_{X_1\geq T_1,X_2<T_2}P_{Y|F} + P_{X_1\geq T_1}P_{t_{x_1}<X_1<T_1,X_2\geq T_2}P_{Y|F}.\end{aligned}$$

Adding $_{X_1\geq T_1}P_{X_1\geq T_1,X_2\geq T_2}P_{Y|T}$ and $P_{X_1} \geq t_{X_1}P_{X_1\geq T_1,X_2<T_2}P_{Y|F}$ to both sides, then dividing by $P_{X_1\geq T_1}P_{X_1\geq t_{X_1}}P_{Y|F}$ yields

$$\begin{aligned}&\frac{P_{X_1\geq T_1,X_2\geq T_2}P_{Y|T} + P_{X_1\geq T_1,X_2<T_2}P_{Y|F}}{P_{X_1\geq T_1}P_{Y|F}} > \\ &\frac{X_1 \geq T_1, X_2 \geq T_2 P_{Y|T} + P_{X_1\geq T_1,X_2<T_2}P_{Y|F} + P_{t_{x_1}<X_1<T_1,X_2\geq T_2}P_{Y|F}}{P_{X_1\geq t_{X_1}}P_{Y|F}}\end{aligned}$$

Which results in $\frac{a_T(c_T+d_T)}{(a_T+b_T)c_T} > \frac{a_t(c_t+d_t)}{(a_t+b_t)c_t}$ for $t < T$. A similar proof follows for $t > T$

B.1.2 *Youden's* First we prove the theorem for $t_{x_1} < T$ and begin with the statement $P_{Y|T} >$

$P_{Y|F}$. Adding $P_{X_1 \geq T_1, X_2 \geq T_2}$ and replacing $P_{X_1 \geq T_1, X_2 \geq T_2}$ with $(1-(P_{X_1 \geq T_1, X_2 < T_2} + P_{X_1 < T_1, X_2 \geq T_2} + P_{X_1 < T_1, X_2 < T_2}))$ pn the right yields

$$P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} > (1 - (P_{X_1 \geq T_1, X_2 < T_2} + P_{X_1 < T_1, X_2 \geq T_2} + P_{X_1 < T_1, X_2 < T_2}))P_{Y|F}$$

Distributing and adding $(P_{X_1 \geq T_1, X_2 < T_2} + P_{X_1 < T_1, X_2 \geq T_2} + P_{X_1 < T_1, X_2 < T_2})P_{Y|F}$ yields,

$$P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} + (P_{X_1 \geq T_1, X_2 < T_2} + P_{X_1 < T_1, X_2 \geq T_2} + P_{X_1 < T_1, X_2 < T_2})P_{Y|F} > P_{Y|F}$$

Adding $P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} P_{Y|F}$ and $-(P_{X_1 \geq T_1, X_2 < T_2} + P_{X_1 < T_1, X_2 \geq T_2} + P_{X_1 < T_1, X_2 < T_2})P_{Y|F}^2$ to both sides and factoring yields,

$$P_{Y=1}(1 - P_{Y|F}) > (1 - P_{Y=1})P_{Y|F}.$$

Distributing and multiplying through by $P_{X_1 < T_1} - P_{X_1 < t_{X_1}}$ yields

$$P_{Y=1}P_{X_1 < T_1} - P_{Y=1}P_{X_1 < T_1}P_{Y|F} >$$
$$(P_{X_1 < T_1} - P_{X_1 < t_{X_1}})P_{Y|F} - P_{Y=1}(P_{X_1 < T_1} - P_{X_1 < t_{X_1}})P_{Y|F} + P_{Y=1}P_{X_1 < t_{X_1}} - P_{Y=1}P_{X_1 < t_{X_1}}P_{Y|F}.$$

Adding $P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} + P_{X_1 \geq T_1, X_2 < T_2} P_{Y|F}$ to both sides and factoring yields

$$(1 - P_{Y=1})(P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} + P_{X_1 \geq T_1, X_2 < T_2} P_{Y|F}) + P_{Y=1}P_{X_1 < T_1}(1 - P_{Y|F}) >$$
$$(1 - P_{Y=1})(P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} + (P_{X_1 \geq T_1, X_2 < T_2} + (P_{X_1 < T_1} - P_{X_1 < t_{X_1}}))P_{Y|F}) + P_{Y=1}P_{X_1 < t_{X_1}}(1 - P_{Y|F})$$

Divide both sides by $P_{Y=1}(1 - P_{Y=1})$ and subtract 1 to yield

$$\frac{(P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} + P_{X_1 \geq T_1, X_2 < T_2} P_{Y|F})}{P_{Y=1}} + \frac{P_{X_1 < T_1}(1 - P_{Y|F})}{(1 - P_{Y=1})} >$$
$$\frac{(P_{X_1 \geq T_1, X_2 \geq T_2} P_{Y|T} + (P_{X_1 \geq T_1, X_2 < T_2} + (P_{X_1 < T_1} - P_{X_1 < t_{X_1}}))P_{Y|F})}{P_{Y=1}} + \frac{P_{X_1 < t_{X_1}}(1 - P_{Y|F})}{(1 - P_{Y=1})}.$$

Which results in $\frac{a_T}{a_T+c_T} + \frac{d_T}{b_T+d_T} - 1 > \frac{a_t}{a_t+c_t} + \frac{d_t}{b_t+d_t} - 1$ for $t < T$. A similar proof follows for $t > T$.

B.1.3 *Chi-Square*

First we prove the theorem for $t_{x_1} < T$ and begin with the statement $P_{X_1 \geq t_{X_1}} > P_{X_1 \geq T_1}$. Multiplying the right hand side by 1 in the form of $P_{X_1 < t_{X_1}} + P_{X_1 \geq t_{X_1}}$ and distributing yields

$$P_{X_1 \geq t_{X_1}} > P_{X_1 \geq T_1} P_{X_1 < t_{X_1}} + P_{X_1|T_1} P_{X_1 > t_{X_1}}.$$

Subtracting $P_{X_1|T_1} P_{X_1 > t_{X_1}}$ from both sides and, factoring and replacing $P_{X_1 > T_1}$ with $P_{X_1 \geq T_1, X_2 \geq T_2} + P_{X_1 \geq T_1, X_2 < T_2}$ yields

$$P_{X_1|T_1} P_{X_1 > t_{X_1}} P_{X_1 < T_1} > P_{X_1 \geq T_1, X_2 \geq T_2} P_{X_1 < t_{X_1}} + P_{X_1 \geq T_1, X_2 < T_2} P_{X_1 < t_{X_1}}.$$

Factoring and replacing $P_{X_1 > T_1}$ with $P_{X_1 \geq T_1, X_2 \geq T_2} + P_{X_1 \geq T_1, X_2 < T_2}$ again yields

$$P_{X_1 \geq t_{X_1}} P_{X_1 < T_1} > P_{X_1 > T_1} P_{X_1 < t_{X_1}}.$$

Multiplying both sides by $\frac{P^2_{X_1 \geq T_1, X_2 \geq T_2} (P_{Y|T} - P_{Y|F})^2}{(1 - P_{Y=1}) P_{Y=1} P_{X_1 \geq T_1} P_{X_1 \geq t_{X_1}}}$ yields

$$\frac{P_{X_1 < T_1} P^2_{X_1 \geq T_1, X_2 \geq T_2} (P_{Y|T} - P_{Y|F})^2}{P_{X_1 \geq T_1} (1 - P_{Y=1}) P_{Y=1}} > \frac{P_{X_1 < t_{X_1}} P^2_{X_1 \geq T_1, X_2 \geq T_2} (P_{Y|T} - P_{Y|F})^2}{P_{X_1 \geq t_{X_1}} (1 - P_{Y=1}) P_{Y=1}}$$

Which can be manipulated into $\frac{(a_T d_T - b_T c_T)^2}{(a_T+b_T)(c_T+d_T)(b_T+d_T)(a_T+c_T)} > \frac{(a_t d_t - b_t c_t)^2}{(a_t+b_t)(c_t+d_t)(b_t+d_t)(a_t+c_t)}$. A similar proof follows for $t > T$.

Theorem 3.2 *For continuous variables X_1 and X_2 and a dichotomous variable Y with prevalence $P_{Y=1}$ and thresholds T_1 and T_2 such that $P_{Y|T} > P_{Y|F}$ (Equation 3.1), the rate of convergence to T_1 is faster when jointly thresholding compared to single thresholding. That is,*

$$\frac{\partial g(T_i|X_1, X_2)}{\partial T_i} > \frac{\partial g(T_i|X_i)}{\partial T_i}$$

for either $i = 1 or\ 2$ when g is one of the six statistics defined in Table 3.3.

To prove this theorem, we first prove the following lemma.

Lemma 3.1 *For continuous variables X_1 and X_2 and a dichotomous variable Y with prevalence $P_{Y=1}$, and thresholds T_1 and T_2 if $P_{Y|T} > P_{Y|F}$ then for functions g defined earlier, $g(T|X_1, X_2) > g(T|X_1)$ where $g(T|X_1, X_2)$ is defined using the joint events (a_J, b_J, c_J, d_J) and $g(T|X_1)$ uses the marginal events (a_S, b_S, c_S, d_S). We conjecture that this Lemma will extend to the case of p continuous variables where the p variables are associated with dichotomous outcome Y through their interaction. This proof can be shown through induction.*

B.2.1 *Relative Risk*

Proof:

We begin with the given statement

$$P_{Y|T} > P_{Y|F}$$

Multiply both sides by $P_{X_1>T_1, X_2<T_2}$

$$P_{Y|T} P_{X_1>T_1, X_2<T_2} > P_{Y|F} P_{X_1>T_1, X_2<T_2}$$

Note $P_{X_1>T_1, X_2<T_2} = P_{X_1>T_1} - P_{X_1>T_1, X_2>T_2}$

$$P_{Y|T}(P_{X_1>T_1} - P_{X_1>T_1, X_2>T_2}) > P_{X_1>T_1, X_2<T_2} P_{Y|F}$$

Distribute

$$P_{Y|T} P_{X_1>T_1} - P_{X_1>T_1, X_2>T_2} P_{Y|T} > P_{X_1>T_1, X_2<T_2} P_{Y|F}$$

Add $P_{X_1>T_1, X_2>T_2} P_{Y|T}$ to both sides

$$P_{Y|T} P_{X_1>T_1} > P_{X_1>T_1, X_2>T_2} P_{Y|T} + P_{X_1>T_1, X_2<T_2} P_{Y|F}$$

Divide both sides by $P_{X_1>T_1}$

$$P_{Y|T} > \frac{P_{X_1>T_1,X_2>T_2}P_{Y|T} + P_{X_1>T_1,X_2<T_2}P_{Y|F}}{P_{X_1>T_1}}$$

Divide both sides by $P_{Y|F}$

$$\frac{P_{Y|T}}{P_{Y|F}} > \frac{P_{X_1>T_1,X_2>T_2}P_{Y|T} + P_{X_1>T_1,X_2<T_2}P_{Y|F}}{P_{X_1>T_1}P_{Y|F}}$$

Multiply left hand side by $\frac{P_{X_1>T_1,X_2>T_2}(P_{X_1>T_1,X_2<T_2}+P_{X_1<T_1,X_2>T_2}+P_{X_1<T_1,X_2<T_2})}{P_{X_1>T_1,X_2>T_2}(P_{X_1>T_1,X_2<T_2}+P_{X_1<T_1,X_2>T_2}+P_{X_1<T_1,X_2<T_2})}$

$$\frac{P_{X_1>T_1,X_2>T_2}P_{Y|T}}{P_{X_1>T_1,X_2>T_2}} \frac{(P_{X_1>T_1,X_2<T_2} + P_{X_1<T_1,X_2>T_2} + P_{X_1<T_1,X_2<T_2})}{(P_{X_1>T_1,X_2<T_2} + P_{X_1<T_1,X_2>T_2} + P_{X_1<T_1,X_2<T_2})P_{Y|F}} > \frac{P_{X_1>T_1,X_2>T_2}P_{Y|T} + P_{X_1>T_1,X_2<T_2}P_{Y|F}}{P_{X_1>T_1}P_{Y|F}}$$

Multiply right hand side by $\frac{P_{X_1<T_1,X_2>T_2}+P_{X_1<T_1,X_2<T_2}}{P_{X_1<T_1,X_2>T_2}+P_{X_1<T_1,X_2<T_2}}$

$$\frac{P_{X_1>T_1,X_2>T_2}P_{Y|T}}{P_{X_1>T_1,X_2>T_2}} \frac{(P_{X_1>T_1,X_2<T_2} + P_{X_1<T_1,X_2>T_2} + P_{X_1<T_1,X_2<T_2})}{(P_{X_1>T_1,X_2<T_2} + P_{X_1<T_1,X_2>T_2} + P_{X_1<T_1,X_2<T_2})P_{Y|F}} > \frac{P_{X_1>T_1,X_2>T_2}P_{Y|T} + P_{X_1>T_1,X_2<T_2}P_{Y|F}}{P_{X_1>T_1}} \frac{P_{X_1<T_1,X_2>T_2} + P_{X_1<T_1,X_2<T_2}}{P_{X_1<T_1,X_2>T_2} + P_{X_1<T_1,X_2<T_2}P_{Y|F}}$$

Note $(a_J + b_J) = P_{X_1>T_1,X_2>T_2}$ and $(a_S + bS) = P_{X_1>T_1}$ Also, $(c_J + d_J) = (P_{X_1>T_1,X_2<T_2} + P_{X_1<T_1,X_2>T_2} + P_{X_1<T_1,X_2<T_2})$ and $(c_S + d_S) = (P_{X_1<T_1,X_2>T_2} + P_{X_1<T_1,X_2<T_2})$. Thus we have,

$$\frac{a_J}{(a_J + b_J)} \frac{(c_J + d_J)}{c_J} > \frac{a_S}{(a_S + b_S)} \frac{(c_S + d_S)}{c_S}$$

Which means

$$RR_J > RR_S$$

B.2.2 *Youden's*

Proof:

We begin with the statement

$$P_{Y=1} > P_{Y|F}$$

This is true because we are given $P_{Y|} > P_{Y|F}$ and $P_{Y=1} = P_{Y|T} + P_{Y|F}$. Next we multiply both sides by $P_{X_1>T_1,X_2<T_2}$

$$P_{Y=1}P_{X_1>T_1,X_2<T_2} > P_{X_1>T_1,X_2<T_2}P_{Y|F}$$

Subtract the terms $P_{Y=1}(P_{Y=1|X_1>T_1,X_2<T_2} + P_{Y=1|X_1<T_1,X_2>T_2} + P_{Y=1|X_1<T_1,X_2<T_2})P_{Y|F}$, $P_{Y=1}(P_{X_1<T_1,X_2>T_2} + P_{X_1<T_1,X_2<T_2})$ and $P_{X_1>t_1,X_2>t_2}P_{Y|T}(1 - P_{Y=1})$ from both sides

$$\begin{aligned}
&P_{X_1>t_1,X_2>t_2}P_{Y|T} - P_{Y=1}P_{X_1>t_1,X_2>t_2}P_{Y|T} + P_{Y=1}P_{X_1>t_1,X_2<t_2} + \\
&P_{Y=1}P_{X_1<t_1,X_2>t_2} + P_{Y=1}P_{X_1<t_1,X_2<t_2} - \\
&P_{Y=1}P_{X_1>t_1,X_2<t_2}P_{Y|F} - P_{Y=1}P_{X_1<t_1,X_2>t_2}P_{Y|F} - P_{Y=1}P_{X_1<t_1,X_2<t_2}P_{Y|F} \\
&> P_{X_1>t_1,X_2>t_2}P_{Y|T} + P_{X_1>T_1,X_2<T_2}P_{Y|F} - P_{Y=1}P_{X_1>t_1,X_2>t_2}P_{Y|T} - \\
&P_{Y=1}P_{X_1>t_1,X_2<t_2}P_{Y|F} + P_{Y=1}P_{X_1<t_1,X_2>t_2} + P_{Y=1}P_{X_1<t_1,X_2<t_2} - \\
&P_{Y=1}P_{X_1<t_1,X_2>t_2}P_{Y|F} - P_{Y=1}P_{X_1<t_1,X_2<t_2}P_{Y|F}
\end{aligned}$$

Combine like terms

$$\begin{aligned}
&(1 - P_{Y=1})(P_{X_1>t_1,X_2>t_2}P_{Y|T}) + P_{Y=1}(P_{Y=1|X_1>T_1,X_2<T_2} + P_{Y=1|X_1<T_1,X_2>T_2} + P_{Y=1|X_1<T_1,X_2<T_2})(1 - P_{Y|F}) > \\
&(1 - P_{Y=1})(P_{X_1>t_1,X_2>t_2}P_{Y|T} + P_{X_1>t_1,X_2<t_2}P_{Y|F}) + P_{Y=1}(P_{X_1<t_1,X_2>t_2} + P_{X_1<t_1,X_2<t_2})(1 - P_{Y|F})
\end{aligned}$$

Divide both sides by $P_{Y=1}(1 - P_{Y=1})$

$$\begin{aligned}
&\frac{P_{X_1>t_1,X_2>t_2}P_{Y|T}}{P_{Y=1}} + \frac{(P_{Y=1|X_1>T_1,X_2<T_2} + P_{Y=1|X_1<T_1,X_2>T_2} + P_{Y=1|X_1<T_1,X_2<T_2})(1 - P_{Y|F})}{(1 - P_{Y=1})} > \\
&\frac{(P_{X_1>t_1,X_2>t_2}P_{Y|T} + P_{X_1>t_1,X_2<t_2}P_{Y|F}}{P_{Y=1}} + \frac{P_{X_1<t_1,X_2>t_2} + P_{X_1<t_1,X_2<t_2})(1 - P_{Y|F})}{(1 - P_{Y=1})}
\end{aligned}$$

Subtract 1 from both sides

$$\frac{P_{X_1>t_1,X_2>t_2}P_{Y|T}}{P_{Y=1}} + \frac{(P_{Y=1|X_1>T_1,X_2<T_2} + P_{Y=1|X_1<T_1,X_2>T_2} + P_{Y=1|X_1<T_1,X_2<T_2})(1-P_{Y|F})}{(1-P_{Y=1})} - 1 >$$
$$\frac{(P_{X_1>t_1,X_2>t_2}P_{Y|T} + P_{X_1>t_1,X_2<t_2}P_{Y|F}}{P_{Y=1}} + \frac{P_{X_1<t_1,X_2>t_2} + P_{X_1<t_1,X_2<t_2})(1-P_{Y|F})}{(1-P_{Y=1})} - 1$$

We have

$$\frac{a_J}{(a_J+c_J)} + \frac{d_J}{(b_J+d_J)} - 1 > \frac{a_S}{(a_S+c_S)} + \frac{d_S}{(b_S+d_S)} - 1$$

Thus

$$Youden's_J > Youden's_S$$

B.2.3 *Gini Index*

Proof:

We begin with the given statement

$$P_{Y|F} < P_{Y|T}$$

Subtract $P_{Y|F}$ from both sides

$$0 < P_{Y|T} - P_{Y|F}$$

Square both sides

$$0 < (P_{Y|T} - P_{Y|F})^2$$

Multiply out the right hand side

$$0 < P_{Y|T}^2 - 2P_{Y|T}P_{Y|F} + P_{Y|F}^2$$

Subtract squared terms from both sides

$$-P_{Y|T}^2 - P_{Y|F}^2 < -2P_{Y|T}P_{Y|F}$$

Multiply both sides by $P_{X_1>T_1,X_2<T_2}P_{X_1>T_1,X_2>T_2}$

$$-P_{X_1>T_1,X_2<T_2}P_{X_1>T_1,X_2>T_2}P_{Y|T}^2 - P_{X_1>T_1,X_2<T_2}P_{X_1>T_1,X_2>T_2}P_{Y|F}^2 < -2P_{X_1>T_1,X_2<T_2}P_{X_1>T_1,X_2>T_2}P_{Y|T}P_{Y|F}$$

Note $P_{X_1>T_1} = P_{X_1>T_1,X_2>T_2} + P_{X_1>T_1,X_2<T_2}$ thus,

$$\begin{aligned}&-(P_{X_1>T_1,X_2>T_2} - P_{X_1>T_1})P_{X_1>T_1,X_2>T_2}P_{Y|T}^2 - P_{X_1>T_1,X_2<T_2}(P_{X_1>T_1,X_2<T_2} - P_{X_1>T_1})P_{Y|F}^2\\&< -2P_{X_1>T_1,X_2<T_2}P_{X_1>T_1,X_2>T_2}P_{Y|T}P_{Y|F}\end{aligned}$$

Distribute

$$\begin{aligned}&-P_{X_1>T_1}P_{X_1>T_1,X_2>T_2}P_{Y|T}^2 - P_{X_1>T_1}P_{X_1>T_1,X_2<T_2}P_{Y|F}^2\\&< -(P_{X_1>T_1,X_2>T_2}^2P_{Y|T}^2 + P_{X_1>T_1,X_2<T_2}^2P_{Y|F}^2 + 2P_{X_1>T_1,X_2<T_2}P_{X_1>T_1,X_2>T_2}P_{Y|T}P_{Y|F})\end{aligned}$$

Factor the right hand side

$$\begin{aligned}&-P_{X_1>T_1}P_{X_1>T_1,X_2>T_2}P_{Y|T}^2 - P_{X_1>T_1}P_{X_1>T_1,X_2<T_2}P_{Y|F}^2\\&< -(P_{X_1>T_1,X_2>T_2}P_{Y|T} + P_{X_1>T_1,X_2<T_2}P_{Y|F})^2\end{aligned}$$

Add $P_{X_1>T_1}P_{X_1>T_1,X_2>T_2}P_{Y|T}$ and $P_{X_1>T_1}P_{X_1>T_1,X_2<T_2}P_{Y|F}$ to both sides

$$- P_{X_1>T_1} P_{X_1>T_1,X_2>T_2} P_{Y|T}^2 - P_{X_1>T_1} P_{X_1>T_1,X_2<T_2} P_{Y|F}^2 + P_{X_1>T_1} P_{X_1>T_1,X_2>T_2} P_{Y|T} +$$

$$P_{X_1>T_1} P_{X_1>T_1,X_2<T_2} P_{Y|F}$$

$$< P_{X_1>T_1} P_{X_1>T_1,X_2>T_2} P_{Y|T} +$$

$$P_{X_1>T_1} P_{X_1>T_1,X_2<T_2} P_{Y|F} - (P_{X_1>T_1,X_2>T_2} P_{Y|T} + P_{X_1>T_1,X_2<T_2} P_{Y|F})^2$$

Factor

$$P_{X_1>T_1} P_{X_1>T_1,X_2>T_2} P_{Y|T}(1 - P_{Y|T}) + P_{X_1>T_1} P_{X_1>T_1,X_2<T_2} P_{Y|F}(1 - P_{Y|F})$$

$$< P_{X_1>T_1} P_{X_1>T_1,X_2>T_2} P_{Y|T} + P_{X_1>T_1} P_{X_1>T_1,X_2<T_2} P_{Y|F} - (P_{X_1>T_1,X_2>T_2} P_{Y|T} + P_{X_1>T_1,X_2<T_2} P_{Y|F})^2$$

Divide both sides by $P_{X_1>T_1}$

$$P_{X_1>T_1,X_2>T_2} P_{Y|T}(1 - P_{Y|T}) + P_{X_1>T_1,X_2<T_2} P_{Y|F}(1 - P_{Y|F})$$

$$< \frac{(P_{X_1>T_1,X_2>T_2} P_{Y|T} + P_{X_1>T_1,X_2<T_2} P_{Y|F})(P_{X_1>T_1} - (P_{X_1>T_1,X_2>T_2} P_{Y|T} + P_{X_1>T_1,X_2<T_2} P_{Y|F}))}{P_{X_1>T_1}}$$

Add $(P_{X_1<T_1,X_2>T_2} + P_{X_1<T_1,X_2<T_2}) P_{Y|F}(1 - P_{Y|F})$ to both sides

$$P_{X_1>T_1,X_2>T_2} P_{Y|T}(1 - P_{Y|T}) + P_{X_1>T_1,X_2<T_2} + P_{X_1<T_1,X_2>T_2} + P_{X_1<T_1,X_2<T_2} P_{Y|F}(1 - P_{Y|F})$$

$$< \frac{(P_{X_1>T_1,X_2>T_2} P_{Y|T} + P_{X_1>T_1,X_2<T_2} P_{Y|F})(P_{X_1>T_1} - (P_{X_1>T_1,X_2>T_2} P_{Y|T} + P_{X_1>T_1,X_2<T_2} P_{Y|F}))}{P_{X_1>T_1}} +$$

$$(P_{X_1<T_1,X_2>T_2} + P_{X_1<T_1,X_2<T_2}) P_{Y|F}(1 - P_{Y|F})$$

Multiply by $\frac{P_{X_1>T_1,X_2>T_2}}{P_{X_1>T_1,X_2>T_2}}$, $\frac{P_{X_1>T_1,X_2<T_2} + P_{X_1<T_1,X_2>T_2} + P_{X_1<T_1,X_2<T_2}}{P_{X_1>T_1,X_2<T_2} + P_{X_1<T_1,X_2>T_2} + P_{X_1<T_1,X_2<T_2}}$, or $\frac{P_{X_1<T_1,X_2>T_2} + P_{X_1<T_1,X_2<T_2}}{P_{X_1<T_1,X_2>T_2} + P_{X_1<T_1,X_2<T_2}}$

$$\frac{P_{X_1>T_1,X_2>T_2}P_{Y|T}P_{X_1>T_1,X_2>T_2}(1-P_{Y|T})}{P_{X_1>T_1,X_2>T_2}}+$$
$$\frac{P_{X_1>T_1,X_2<T_2}+P_{X_1<T_1,X_2>T_2}+P_{X_1<T_1,X_2<T_2}P_{Y|F}(P_{X_1>T_1,X_2<T_2}+P_{X_1<T_1,X_2>T_2}+P_{X_1<T_1,X_2<T_2})(1-P_{Y|F})}{(P_{X_1>T_1,X_2<T_2}+P_{X_1<T_1,X_2>T_2}+P_{X_1<T_1,X_2<T_2})}$$
$$<\frac{(P_{X_1>T_1,X_2>T_2}P_{Y|T}+P_{X_1>T_1,X_2<T_2}P_{Y|F})(P_{X_1>T_1}-(P_{X_1>T_1,X_2>T_2}P_{Y|T}+P_{X_1>T_1,X_2<T_2}P_{Y|F}))}{P_{X_1>T_1}}+$$
$$\frac{(P_{X_1<T_1,X_2>T_2}+P_{X_1<T_1,X_2<T_2})P_{Y|F}(P_{X_1<T_1,X_2>T_2}+P_{X_1<T_1,X_2<T_2})(1-P_{Y|F})}{(P_{X_1<T_1,X_2>T_2}+P_{X_1<T_1,X_2<T_2})}$$

This is equivalent to

$$(\frac{ab}{(a+b)}+\frac{cd}{(c+d)})_J<(\frac{ab}{(a+b)}+\frac{cd}{(c+d)})_S$$

Multiply by a negative

$$-(\frac{ab}{(a+b)}+\frac{cd}{(c+d)})_J>-(\frac{ab}{(a+b)}+\frac{cd}{(c+d)})_S$$

Thus

$$Gini_J>Gini_S$$

B.2.4 *Chi-Square*

Proof We begin with the true statement $P_{X_1>T_1}=P_{X_1>T_1,X_2>T_2}+P_{X_1>T_1,X_2<T_2}$ This means that

$$P_{X_1>T_1}>P_{X_1>T_1,X_2>T_2}$$

Subtract $P_{X_1>T_1}P_{X_1>T_1,X_2>T_2}$ from both sides

$$P_{X_1>T_1}-P_{X_1>T_1}P_{X_1>T_1,X_2>T_2}>P_{X_1>T_1,X_2>T_2}-P_{X_1>T_1}P_{X_1>T_1,X_2>T_2}$$

Factor

$$(1 - P_{X_1>T_1,X_2>T_2})(P_{X_1>T_1}) > (1 - P_{X_1>T_1})(P_{X_1>T_1,X_2>T_2} >)$$

Multiply both sides by $(P_{Y|T} - P_{Y|F})^2(1 - P_{Y=1})(P_{Y=1})$

$$(1 - P_{X_1>T_1,X_2>T_2})(P_{X_1>T_1})(P_{Y|T} - P_{Y|F})^2(1 - P_{Y=1})(P_{Y=1}) >$$
$$(1 - P_{X_1>T_1})(P_{X_1>T_1,X_2>T_2})(P_{Y|T} - P_{Y|F})^2(1 - P_{Y=1})(P_{Y=1})$$

Divide both sides by $(1 - P_{Y=1})(P_{Y=1})P_{X_1>T_1}$

$$\frac{(P_{Y|T} - P_{Y|F})^2 P_{X_1>T_1,X_2>T_2}(1 - P_{X_1>T_1,X_2>T_2})}{(1 - P_{Y=1})(P_{Y=1})} >$$
$$\frac{(P_{Y|T} - P_{Y|F})^2 (P_{X_1>T_1,X_2>T_2})^2(1 - P_{X_1>T_1})}{(1 - P_{Y=1})(P_{Y=1})P_{X_1>T_1}}$$

Multiplying by a form of 1 on both sides gives

$$\frac{(P_{Y|T} - P_{Y|F})^2 (P_{X_1>T_1,X_2>T_2})^2(1 - P_{X_1>T_1,X_2>T_2})^2}{(1 - P_{Y=1})(P_{Y=1})(P_{X_1>T_1,X_2>T_2})(1 - P_{X_1>T_1,X_2>T_2})} >$$
$$\frac{(P_{Y|T} - P_{Y|F})^2 (P_{X_1>T_1,X_2>T_2})^2(1 - P_{X_1>T_1})^2}{(1 - P_{Y=1})(P_{Y=1})P_{X_1>T_1}(1 - P_{X_1>T_1})}$$

Some sophisticated manipulation leads to

$$(\frac{(ad - bc)^2}{(a + b)(c + d)(b + d)(a + c)})_J > (\frac{(ad - bc)^2}{(a + b)(c + d)(b + d)(a + c)})_S$$

and

$$(CHI)_J > (CHI)_S$$

LIST OF REFERENCES

[1] T.A. Manolio and F.S. Collins. Genes, environment, health, and disease: Facing up to complexity. *Hum Hered*, 63(2):63–66, 2007.

[2] W.Y. Loh. Classification and regression tree methods. *Encyclopedia of statistics in quality and reliability*, 2008.

[3] W.J. Catalona, D.S. Smith, and D.K. Ornstein. Prostate cancer detection in men with serum psa concentrations of 2.6 to 4.0 ng/ml and benign prostate examination: Enhancement of specificity with free psa measurements. *JAMA*, 277(18):1452–1455, 1997.

[4] P.W.F. Wilson, R.B. D'Agostino, D. Levy, A.M. Belanger, H. Silbershatz, and W.B. Kannel. Prediction of coronary heart disease using risk factor categories. *Circulation*, 97(18):1837–1847, 1998.

[5] N.J. Perkins and E.F. Schisterman. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology*, 163(7):670–675, 2006.

[6] I. Ruczinski, C. Kooperberg, and M. LeBlanc. Logic regression. *Journal of Computational and Graphical Statistics*, 12(3):475–511, 2003.

[7] C. Contal and J. O'Quigley. An application of changepoint methods in studying the effect of age on survival in breast cancer. *Computational Statistics Data Analysis*, 30(3):253 – 270, 1999.

[8] M. Mazumdar, A Smith, and J. Bacik. Methods for categorizing a prognostic variable in a multivariable setting. *Statistics in medicine*, 22(4):559–571, 2003.

[9] P. Royston, W. Sauerbrei, and D.G. Altman. Modeling the effects of continuous risk factors. *Journal of clinical epidemiology*, 53(2):219–220, 2000.

[10] P. Royston, D.G. Altman, and W. Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, 25(1):127–141, 2006.

[11] B.A. McKinney, D.M. Reif, M.D. Ritchie, and J.H. Moore. Machine learning for detecting gene-gene interactions. *Applied Bioinformatics*, 5(2):77–88, 2006.

[12] I. Lobo. Epistasis: Gene interaction and the phenotypic expression of complex diseases. *Nature Education*, 2008.

[13] K.M. Uramoto, C. Michet Jr., J. Thumboo, J. Sunku, W.M. O'Fallon, and S.E. Gabriel. Trends in the incidence and mortality of systemic lupus erythematosus, 1950 - 1992. *Arthritis & Rheumatism*, 42(1):46–50, 1999.

[14] J.J. Sacks, C.G. Helmick, G. Langmaid, and J.E. Sniezek. Trends in death from systemic lupus erythematosus. *MMWR Morb Mortal Wkly Rep.*, 51(17):371–4, May 2002.

[15] C.G. Helmick, D.T. Felson, R.C. Lawrence, S. Gabriel, R. Hirsch, C.K. Kwoh, M.H. Liang, H.M. Kremers, M. Mayes, P.A. Merkel, S.R. Pillemer, J.D. Reveille, J.H. Stone, and National Arthritis Data Workgroup. Estimates of the prevalence of arthritis and other rheumatic conditions in the united states: Part i. *Arthritis Rheumatism*, 58(1):15–25, 2008.

[16] Emily C. Somers, Wendy Marder, Patricia Cagnoli, Emily E. Lewis, Peter DeGuire, Caroline Gordon, Charles G. Helmick, Lu Wang, Jeffrey J. Wing, J. Patricia Dhar, James Leisen, Diane Shaltis, and W. Joseph McCune. Population-based incidence and prevalence of systemic lupus erythematosus: The michigan lupus epidemiology and surveillance program. *Arthritis Rheumatology*, 66(2):369–378, 2014.

[17] C. Franco, W Yoo, D. Franco, and Z. Xu. Predictors of end stage renal disease in african americans with lupus nephritis. *Bulletin of the NYU Hospital for Joint Diseases*, (1936-9719 (Linking)):–, 2010.

[18] D. Deafen, A. Escalante, L. Weinrib, D. Horwitz, B. Bachman, P. Roy-Burman, A. Walker, and T.M. Mack. A revised estimate of twin concordance in systemic lupus erythematosus. *Arthritis Rheumatism*, 35(3):311–318, 1992.

[19] S.E. Vaughn, L.C. Kottyan, M.E. Munroe, and J.B. Harley. Genetic susceptibility to lupus: the biological basis of genetic risk found in b cell signaling pathways. *Journal of Leukocyte Biology*, 92(3):577–591, 2012.

[20] G. Zandman-Goddard, M. Solomon, Z. Rosman, and Y. Shoenfeld. Enviornment and lupus-related diseases. *Lupus*, 2014.

[21] JA Sparks and KH Costenbader. Genetics, environment, and gene-environment interactions in the development of systemic rheumatic diseases. *Rheumatic diseases clinics of North America*, 40(4):637–657, 2014.

[22] P. Peduzzi, J. Concato, E. Kemper, T.R. Holford, and A.R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology*, 49(12):1373–1379, 1996.

[23] J.V. Tu. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of clinical epidemiology*, 49(11):1225–1231, 1996.

[24] S. Dreiseitl and L. Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, 35(5âĂŞ6):352 – 359, 2002.

[25] S. Balakrishnama and A. Ganapathiraju. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, 18, 1998.

[26] L. Breiman. Technical note: Some properties of splitting criteria. *Machine Learning*, 24(1):41–47, 1996.

[27] G.G. Moisen. *Classification and regression trees*. Elsevier, 2008.

[28] W. Yoo, B.A. Ference, M. L. Cote, and A. Schwartz. A comparison of logistic regression, logic regression, classification tree, and random forests to identify effective gene-gene and gene-environmental interactions. *International Journal of Applied Science and Technology*, 2(7):268âĂŞ, 2012.

[29] R.C. MacCallum, S. Zhang, K.J. Preacher, and Rucker D.D. On the practice of dichotomization of quantitative variables. *Pyschological Methods*, 7(1):19–40, 2002.

[30] W.B. Kannel and D.L. McGee. Diabetes and glucose tolerance as risk factors for cardiovascular disease: the framingham study. *Diabetes Care*, 2(2):120–126, 1979.

[31] K.K. Ray, J.P. Kastelein, S.M. Boekholdt, S.J. Nicholls, K.T. Khaw, C.M. Ballantyne, A.L. Catapano, Z. Reiner, and T.F. Lüscher. The acc/aha 2013 guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular disease risk in adults: the good the bad and the uncertain: a comparison with esc/eas guidelines for the management of dyslipidaemias 2011. *European heart journal*, page ehu107, 2014.

[32] L. Breiman, J.H.. Friedman, R.A. Olshen, and C.J. Stone. Classification and regression trees. wadsworth & brooks. *Monterey, CA*, 1984.

[33] B.E. Hansen. Sample splitting and threshold estimation. *Econometrica*, 68(3):575–603, 2000.

[34] D.S. Goodman, S.B. Hulley, and L.T. Clark. Report of the national cholesterol education program expert panel on detection, evaluation, and treatment of high blood cholesterol in adults. *Archives of Internal Medicine*, 148(1):36–69, 1988.

[35] DS Goodman. The national cholesterol education program: guidelines, status and issues. *American Journal of Medicine*, 90:32s–35s, 1991.

[36] J. Naggara, O.and Raymond, F. Guilbert, A. Roy, D. Weill, and DG Altman. Analysis by categorizing or dichotomizing continuous variables is inadvisable: An example from the natural history of unruptured aneurysms. *American Journal of Neuroradiology*, 32, 2011.

[37] H. SchÃd'fer. Constructing a cut-off point for a quantitative diagnostic test. *Statistics in Medicine*, 8(11):1381–1391, 1989.

[38] J. Vermont, J.L. Bosson, P. Francois, C. Robert, A. Rueff, and J. Demongeot. Strategies for graphical threshold determination. *Computer Methods and Programs in Biomedicine*, 35:141–150, 1991.

[39] R.J. Gallop, P. Crits-Christoph, L.R. Muenz, and X.M. Tu. Determination and interpretation of the optimal operating point for roc curves derived through generalized linear models. *Understanding Statistics*, 2(4):219–242, 2003.

[40] A.L. Bortheiry, D.A. Malerbi, and L.J. Franco. The roc curve in the evaluation of fasting capillary blood glucose as a screening test for diabetes and igt. *Diabetes Care*, 17:1269âĂŞ1272, 1994.

[41] R.M. Hoffman, D.L. Clanon, B. Littenberg, J.J. Frank, and J.C. Peirce. Using the free-to-total prostate-specific antigen ratio to detect prostate cancer in men with nonspecific elevations of prostate-specific antigen levels. *J. Gen. Intern Med*, 15:739âĂŞ748, 2000.

[42] G. Alvarez-GarcÄśa, E. Collantes-Fernandez, E. Costas, X. Rebordosa, and L. Ortega-Mora. Influence of age and purpose for testing on the cut-off selection of serological methods in bovine neosporosis. *Veterinary Research, BioMed Central*, 34(3):341–352, 2003.

[43] S. Manel, H.C. Williams, and S.J. Ormerod. Evaluating presenceâĂŞabsence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, 38(5):921–931, 2001.

[44] M.J. Kelly, F.D. Dunstan, K. Lloyd, and D. Fone. Evaluating cutpoints for the mhi-5 and mcs using the ghq-12: a comparison of five different methods. *BMC Psychiatry*, 2008.

[45] A.G. Lalkhen and McCluskey A. Clinical tests: sensitivity and specificity. *Continuing Education in Anesthesia, Critical Care and Pain*, 8(6):221–223, 2008.

[46] W.J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950.

[47] H.C. Kraemer. *Risk ratios, odds ratio, and the test QROC. In: Evaluating medical tests*. Newbury Park, CA: SAGE Publications, Inc, 1992.

[48] M. Greiner, D. Pfeiffer, and R.D. Smith. Principals and practical application of the receiver operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine*, 45:23–41, 2000.

[49] D. Boehning, H. Holling, and V. Patilea. A limitation of the diagnostic-odds ratio in determining an optimal cut-off value for a continuous diagnostic test. *Statistical Methods in Medical Research*, 20(5):541–550, 2011.

[50] M. Greiner. Two-graph receiver operating characteristic (tg-roc): a microsoft-excel template for the selection of cut-off values in diagnostic tests. *Journal of Immunological Methods*, 185(1):145–146, 1995-09-11T00:00:00.

[51] C. Strobl, A.L. Boulesteix, and Augustin T. Unbiased split selection for classification trees based on the gini index. *Computational Statistics and Data Analysis*, 52:483–501, 2007.

[52] M. Lopez-Raton, M.X. Rodriguez-Alvarez, C. Cardosa-Suarez, and F. Gude-Sampedro. Optimalcutpoints: An r package for selecting optimal cutpoints in diagnostic testing. *Journal of Statistical Software*, 61(8):1–36, 2014.

[53] K. Aoki, J. Misumi, T. Kimura, W. Zhao, and T. Xie. Evaluation of cutoff levels for screening of gastric cancer using serum pepsinogens and distributions of levels of serum pepsinogen i, ii and of pg i / pg ii ratios in a gastric cancer case-control study. *Journal of Epidemiology*, 7(3):143–151, 1997.

[54] D.J. Hand. Screening vs prevalence estimation. *Applied Statistics*, pages 1–7, 1987.

[55] Institute for Statistics and Mathematics of WirtschaftsuniversitÃd't Wien. The comprehensive r network.

[56] H.K. Tabor, N.J. Risch, and R.M. Myers. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Reviews Genetics*, 3:391–397, 2002.

[57] J.H. Moore, F.W. Asselbergs, and S.M. Williams. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26(4):445–455, 2010.

[58] A. Hughes, J. Chen, M.C. Cornelis, L.B. Chibnik, E.W. Karlson, and Kraf. Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases. *The American Journal of Human Genetics*, 90:962–972, 2012.